Applied Genome Research

DNA isolation

205048 & 205049

Boas Pucker

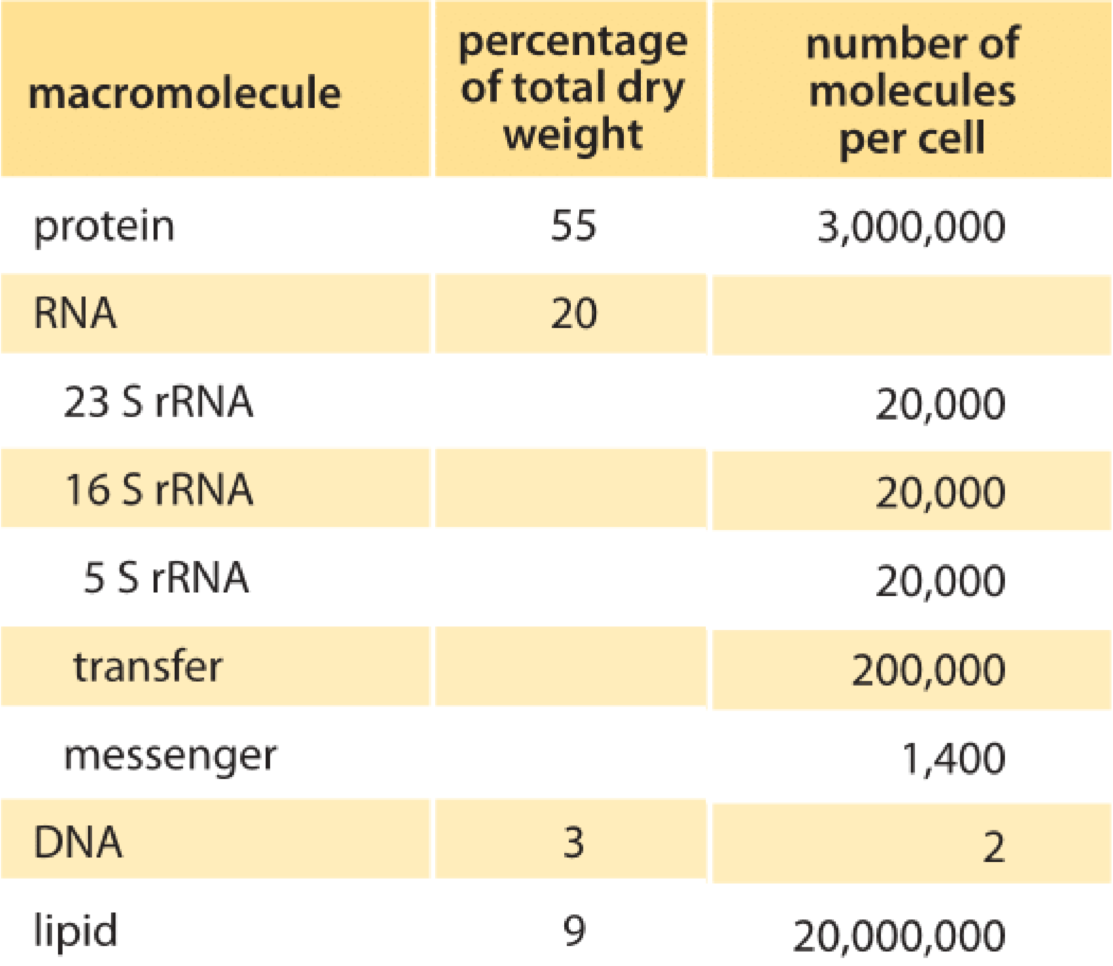
Major types of DNA in eukaryotes (plants)

- gDNA from the nucleus
- mtDNA from the mitochondria (chondrome)
- cpDNA from the chloroplast (plastome)
- pDNA (plasmids)

Problems in DNA isolation for sequencing

- High amount of cpDNA is a big issue in sequencing projects
  - 50-100 chloroplasts per cell with 50-100 plastome per chloroplast
- Sequencing capacity is wasted on the cpDNA molecules
  - Very high sequencing coverage for the plastome, but reduced sequencing coverage for nucleome
- Reducing amount of chloroplasts by incubating plants in the dark for some days prior to DNA isolation
  - Reduced amount of chloroplasts
  - Reduced concentration of starch

… more problems!



macromolecule	percentage of total dry weight	number of molecules per cell
protein	55	3,000,000
RNA	20	
23 S rRNA		20,000
16 S rRNA		20,000
5 S rRNA		20,000
transfer		200,000
messenger		1,400
DNA	3	2
lipid	9	20,000,000

This numbers describe an average *E. coli* cell, but are similar for all other cells

(modified from bionumbers.org)

Established DNA isolation methods

Plant genomic DNA

- Edwards preparation: low quality but quick
- CTAB: high quality but slow
- CARLSON buffer and Genomic-tip: very high quality but slow and expensive

Plasmid DNA

- TELT: cheap and good quality for small plasmids
- Alkaline lysis: cheap and good quality
- Standard plasmid isolation kit: high quality but expensive

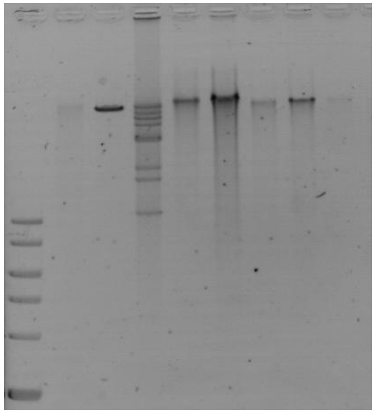
Concept – Plant DNA isolation

- Destruction of cell walls e.g. by grinding
- Resolving powder in lysis buffer
  - Cetyl trimethylammonium bromide (CTAB): solubility of polysaccharides and DNA differs
  - Polyvinyl pyrrolidone (PVP): prevents interaction of phenolic compounds and DNA
  - RNase (A): degrades RNA
  - ß-mercaptoethanol: destroys cell proteins
  - EDTA: protects DNA by inhibiting DNases by capturing Mg^{2+}
- Separation of nucleic acids from other components via chloroform:isoamylalcohol
- Precipitation of DNA for further purification
- Resolving DNA in elution buffer (Tris-HCl)
  - EDTA can be added to protect DNA

Quality Control

- NanoDrop / photometric quantification

- Agarose gel analysis



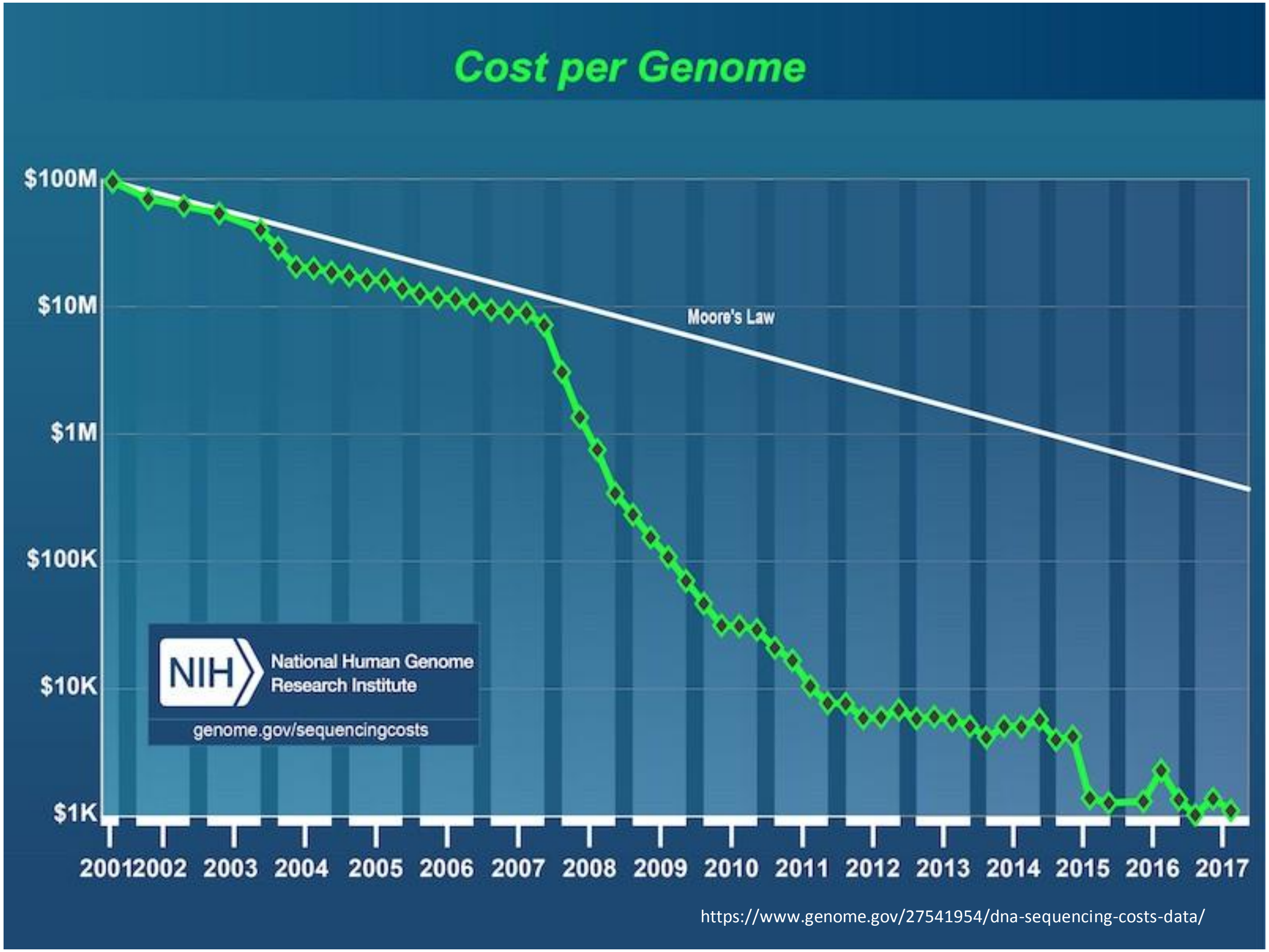
- Agilent chip

Sequencing Technologies - Overview

- First generation:
  - Maxam-Gilbert
  - **Sanger**
- Second generation:
  - SOLiD
  - **Roche 454**
  - **Illumina**
  - IonTorrent
- Third and following generation:
  - SMRT sequencing (PacBio)
  - Oxford Nanopore Technologies (ONT)

...

Sequencing costs / data



https://www.genome.gov/27541954/dna-sequencing-costs-data/

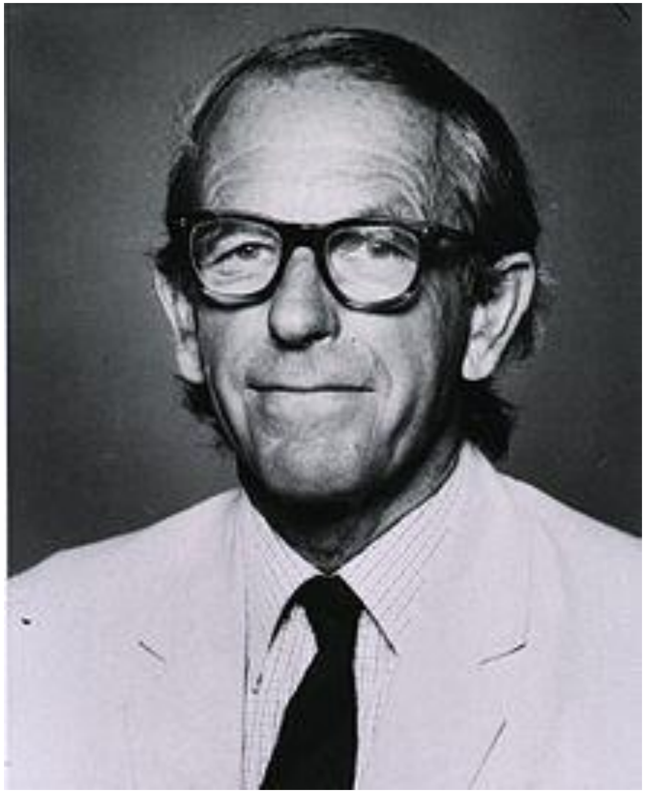
Sanger sequencing

Frederick Sanger

Nobel prices for:

1) Protein sequencing (1958)

2) DNA sequencing (1980)



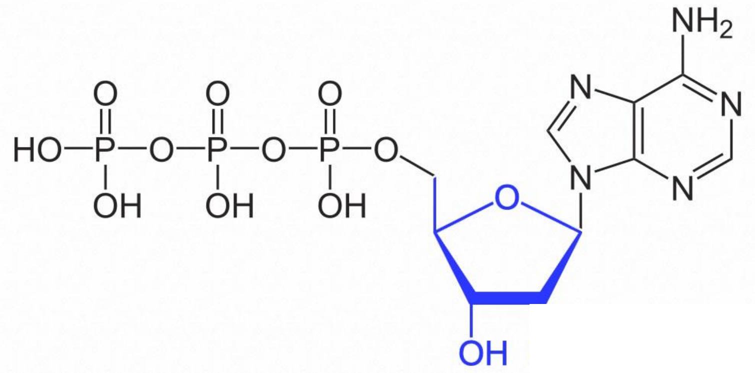
1918-2013

(wikipedia.org)

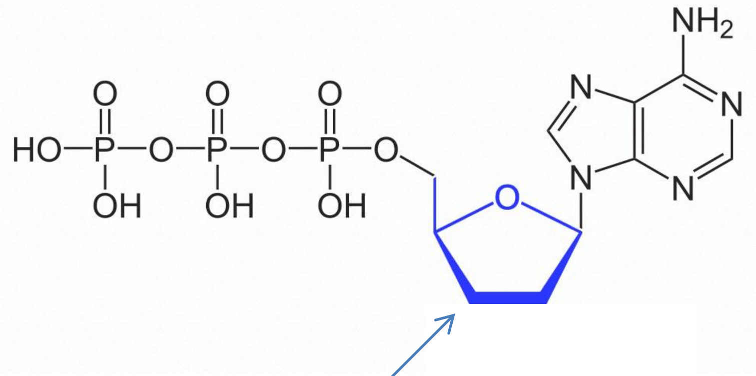
QUESTION

- Which is the direction of biological nucleic acid synthesis?
  - A: 3’ >> 5’
  - B: 5’ >> 3’

Sanger – biochemical concept



dNTP



ddNTP

Absence of 3’-OH group prevents the elongation of the DNA strand.

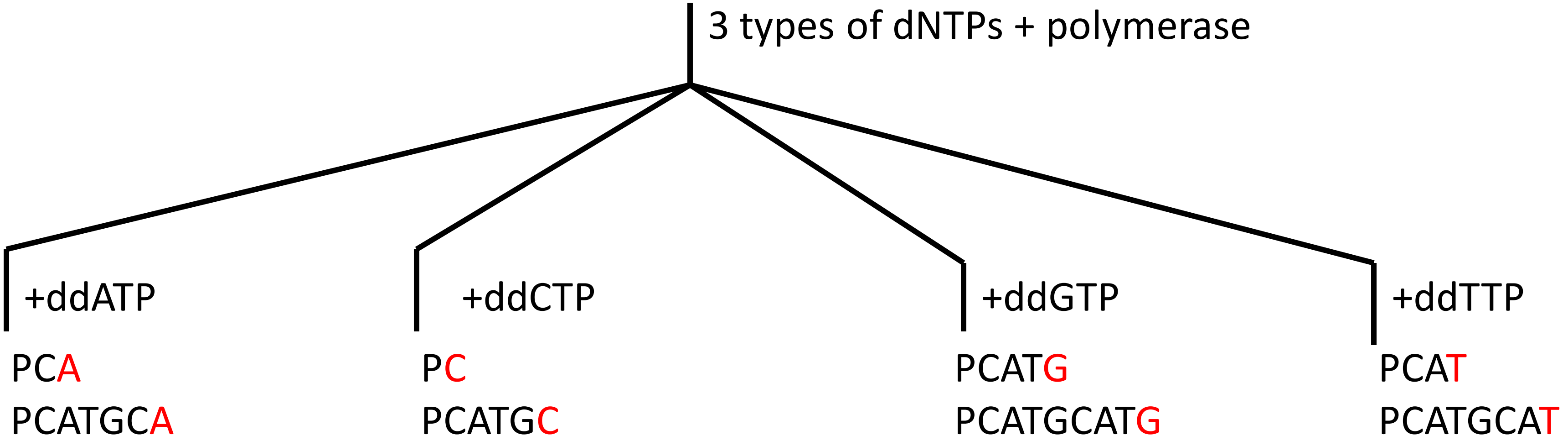
(wikimedia.org)

Biological nucleic acid synthesis: always 5’ >> 3’

Sanger – application concept

Primer (P): 5’ –TGCATGGCATGATGCATG-3’

Template: 3’ –ACGTACCGTACTACGTACGTACGTACGTCTAGGT-5’



3 types of dNTPs + polymerase

+ddATP

PCA

PCATGCA

+ddCTP

PC

PCATGC

+ddGTP

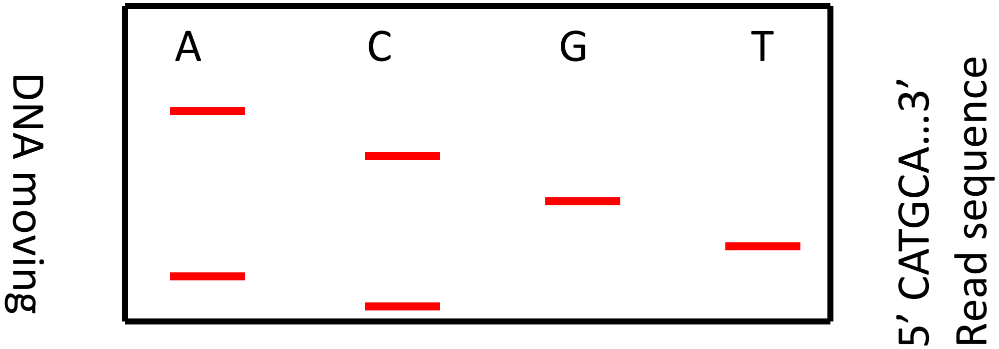
PCATG

PCATGCATG

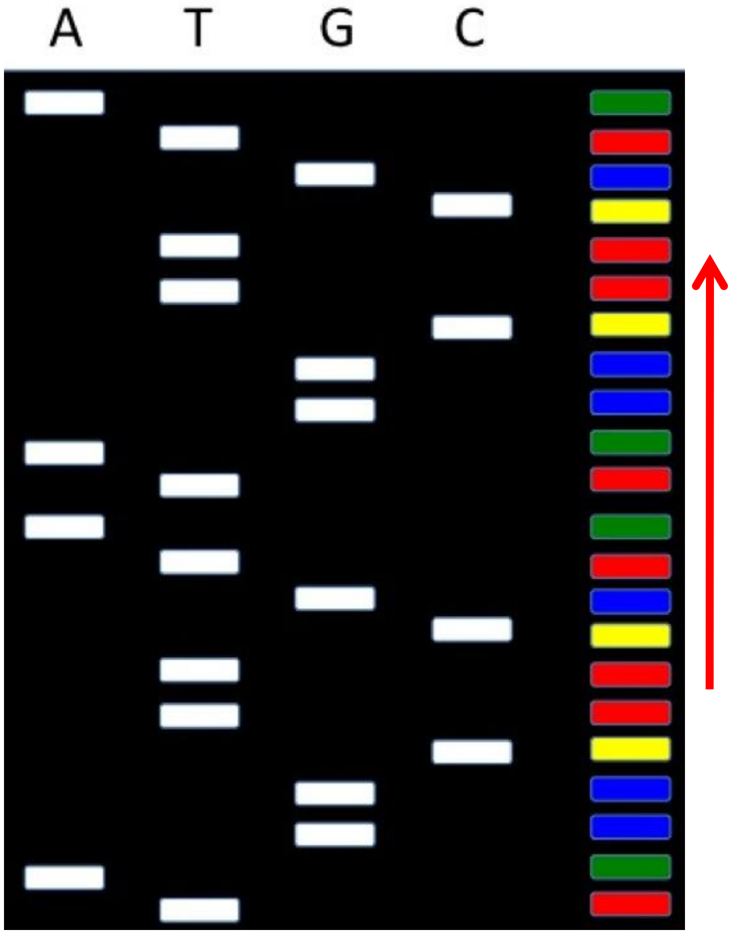
+ddTTP

PCAT

PCATGCAT



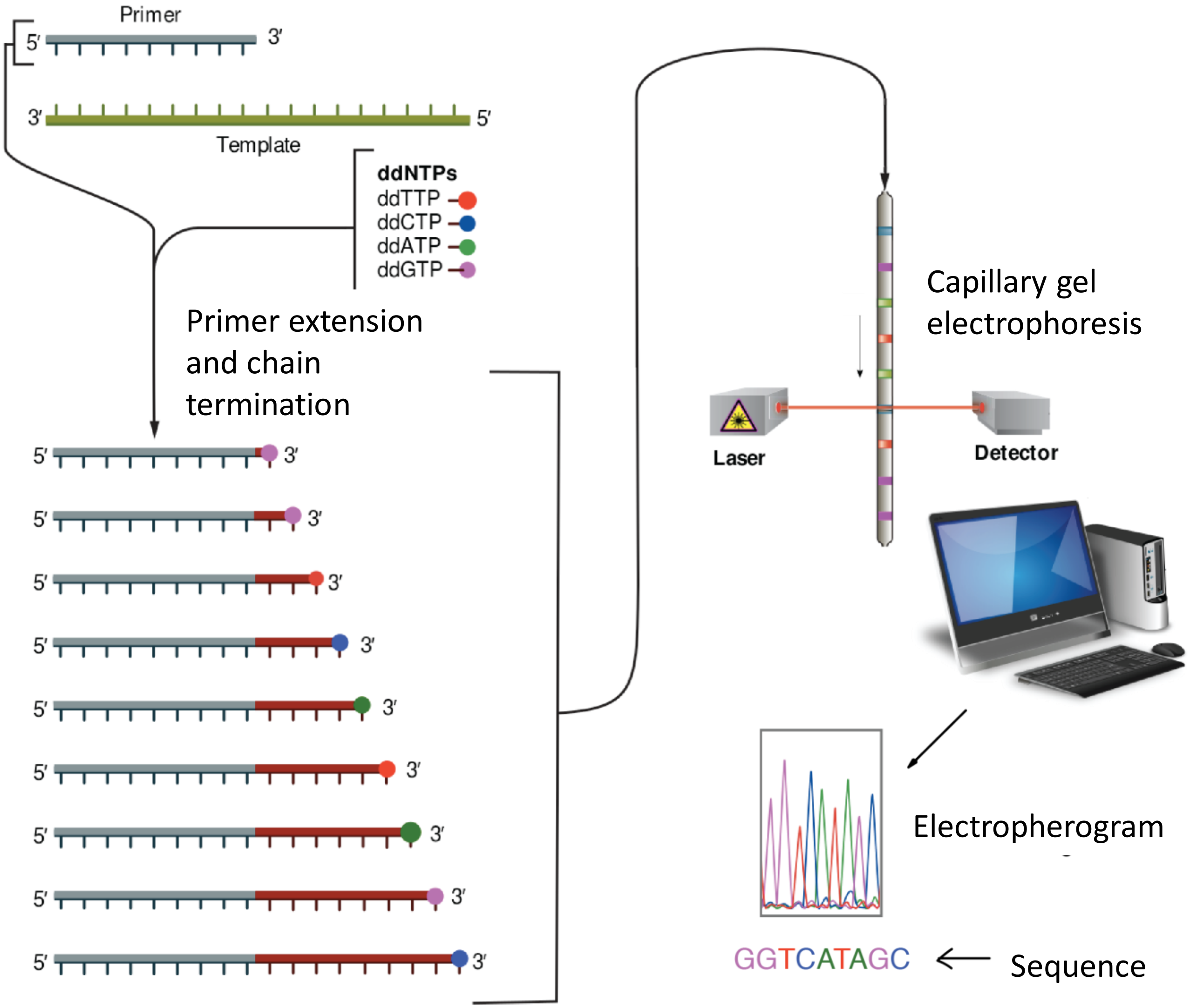
Sanger – original version



(modified from wikimedia.org)

Two persons are analyzing the gel: one is calling the base (‘basecaller’) and the other person is writing down the bases.

Sager – today



Primer extension and chain termination

Capillary gel electrophoresis

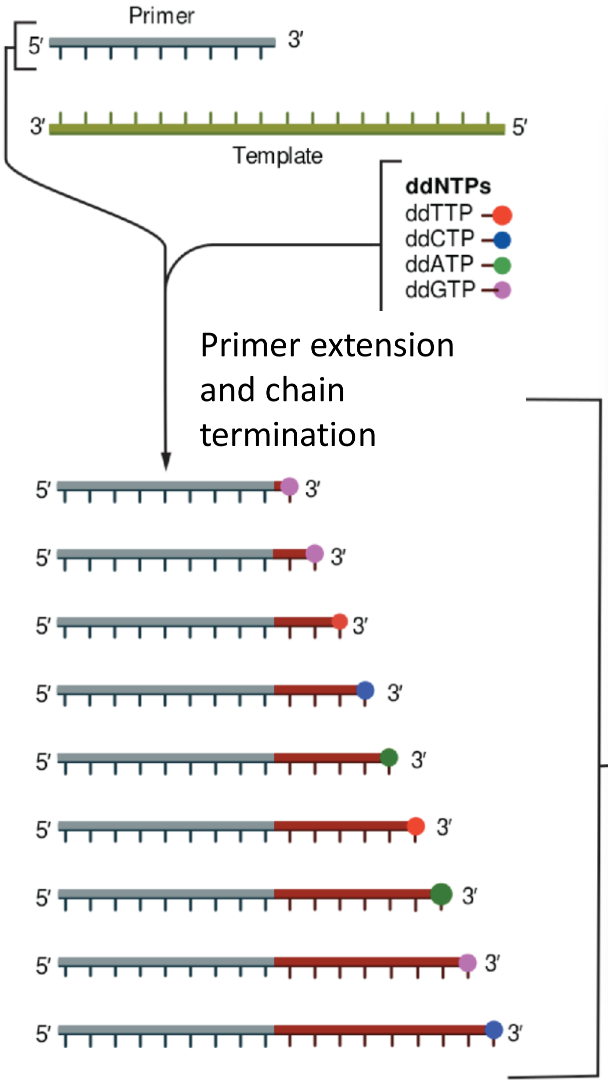
Electropherogram

GGTCATAGC ← Sequence

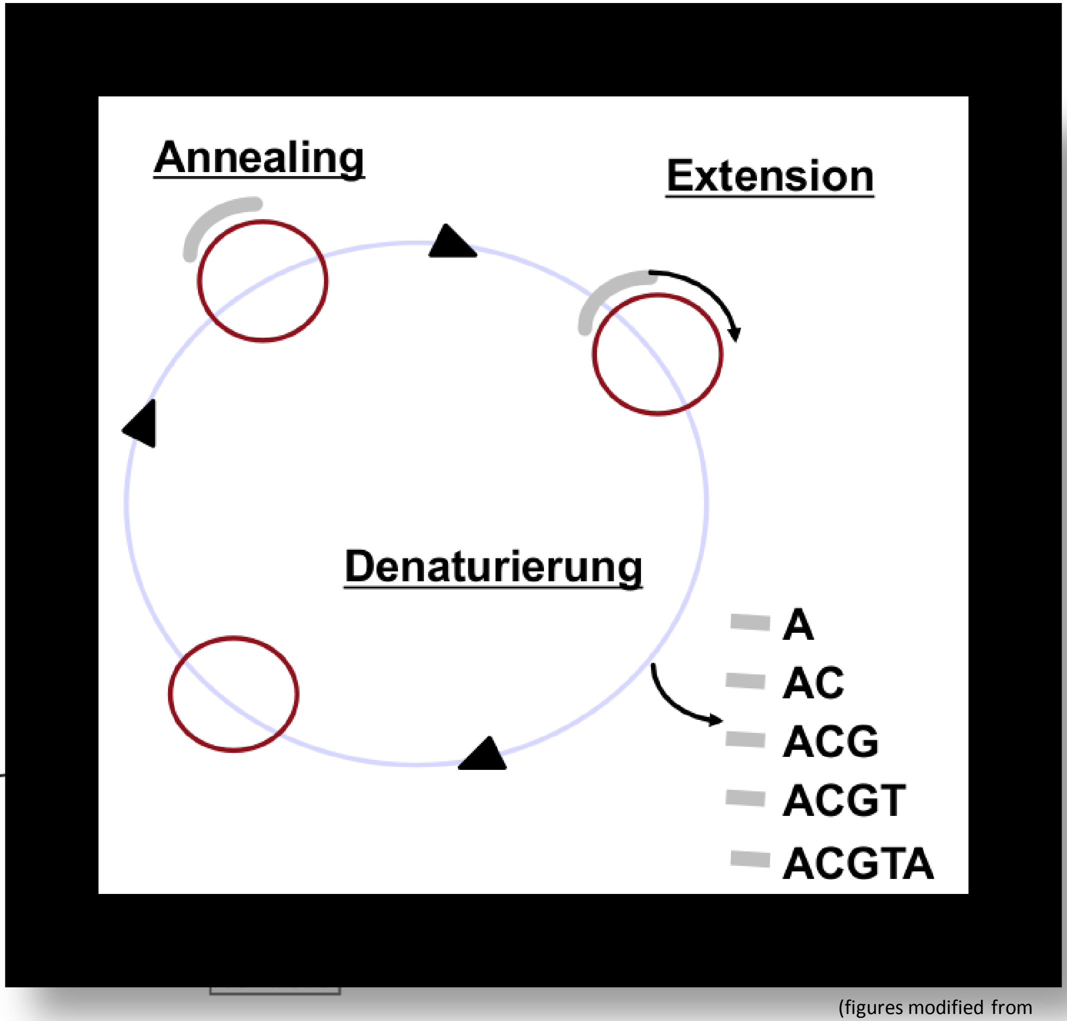
(figures modified from wikipedia.org)

(modified from wikimedia.org)

Sager – today



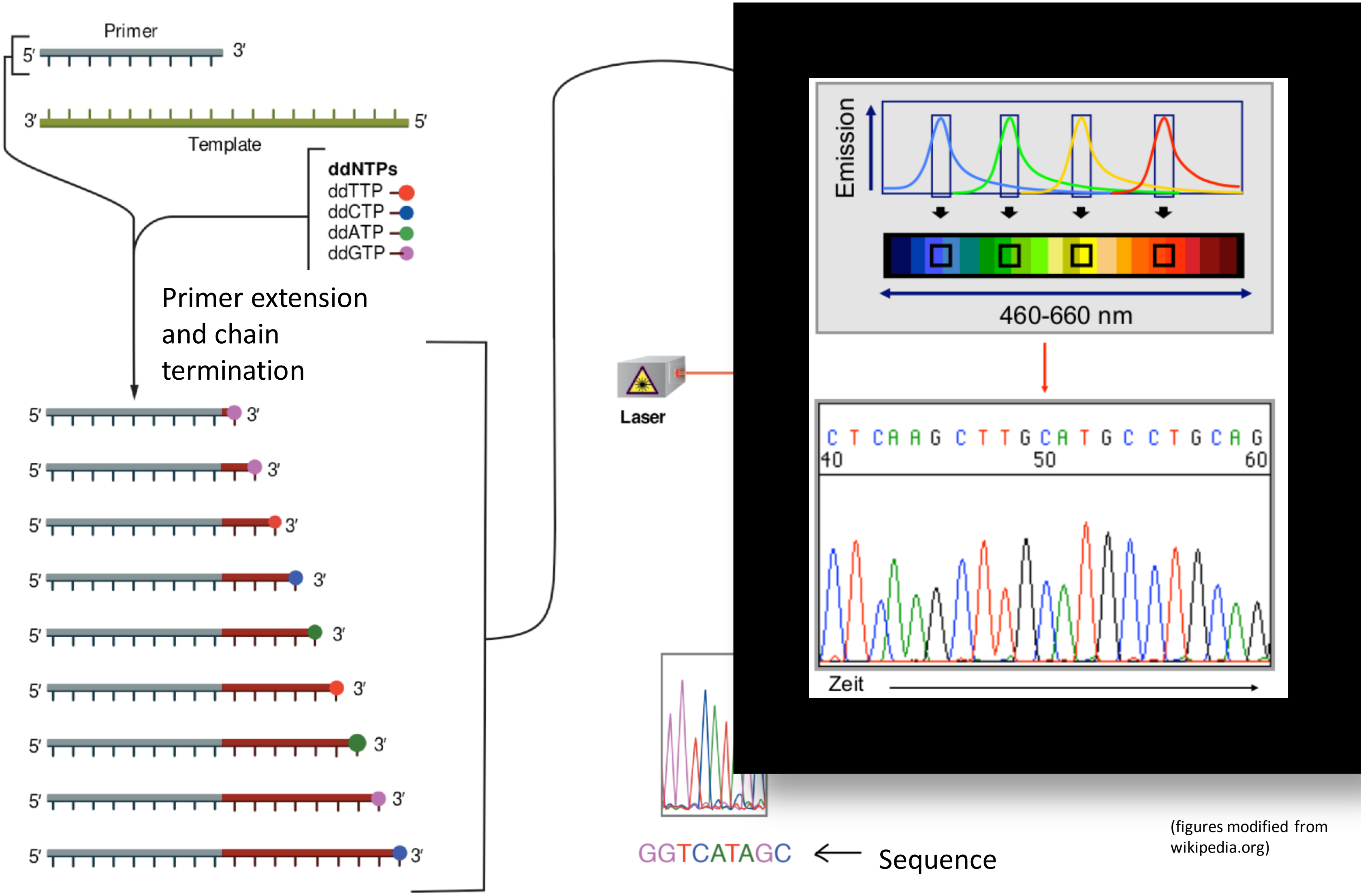
Primer extension and chain termination



GGTCATAGC ← Sequence

(figures modified from wikipedia.org)

Sager – today



Primer extension and chain termination

← Sequence

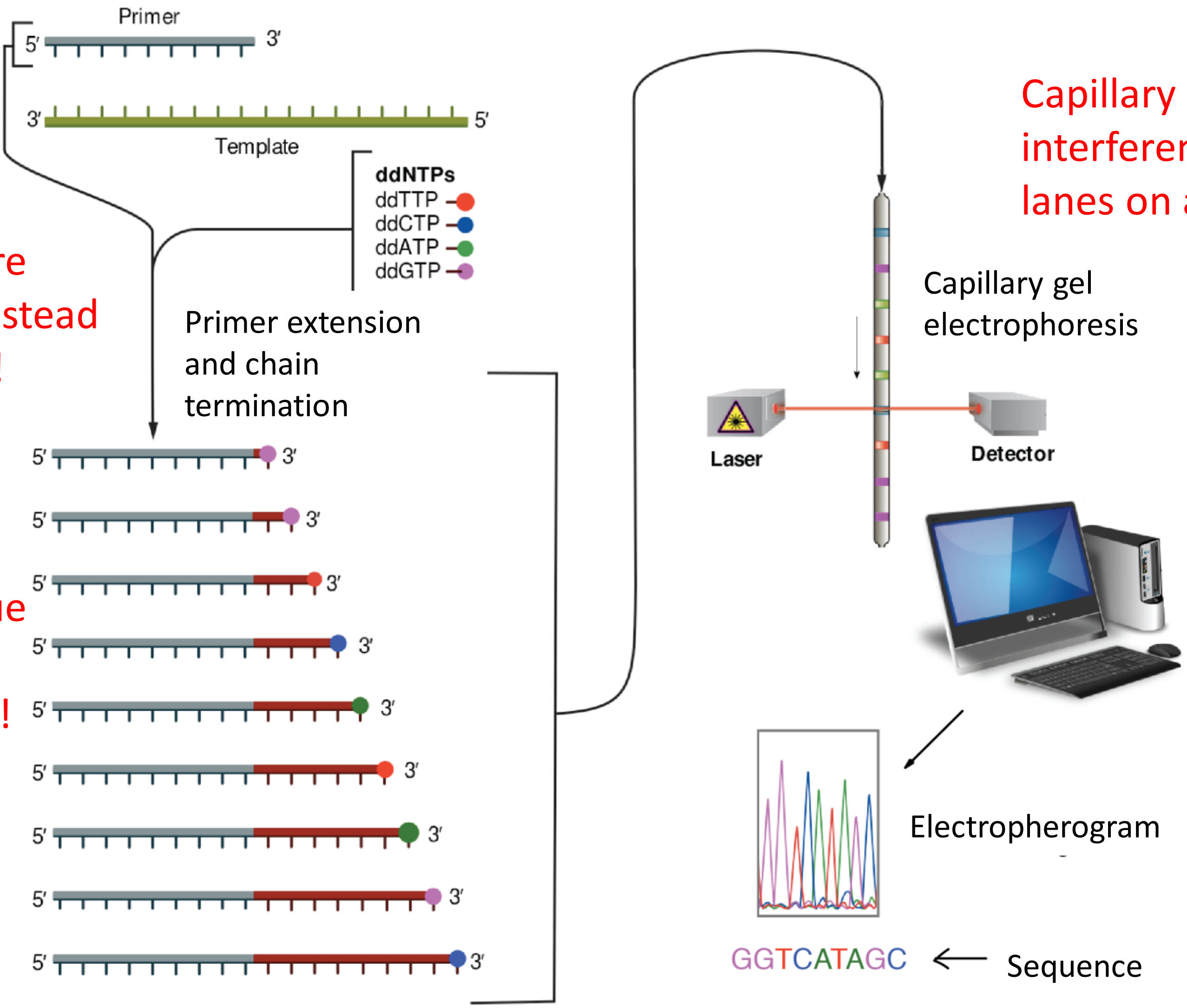
(figures modified from wikipedia.org)

Sager – today

Only one reaction!

ddNTPs are marked instead of primer!

Low input required due to cycle sequencing!



Primer extension and chain termination

Capillary gel electrophoresis

Electropherogram

Sequence

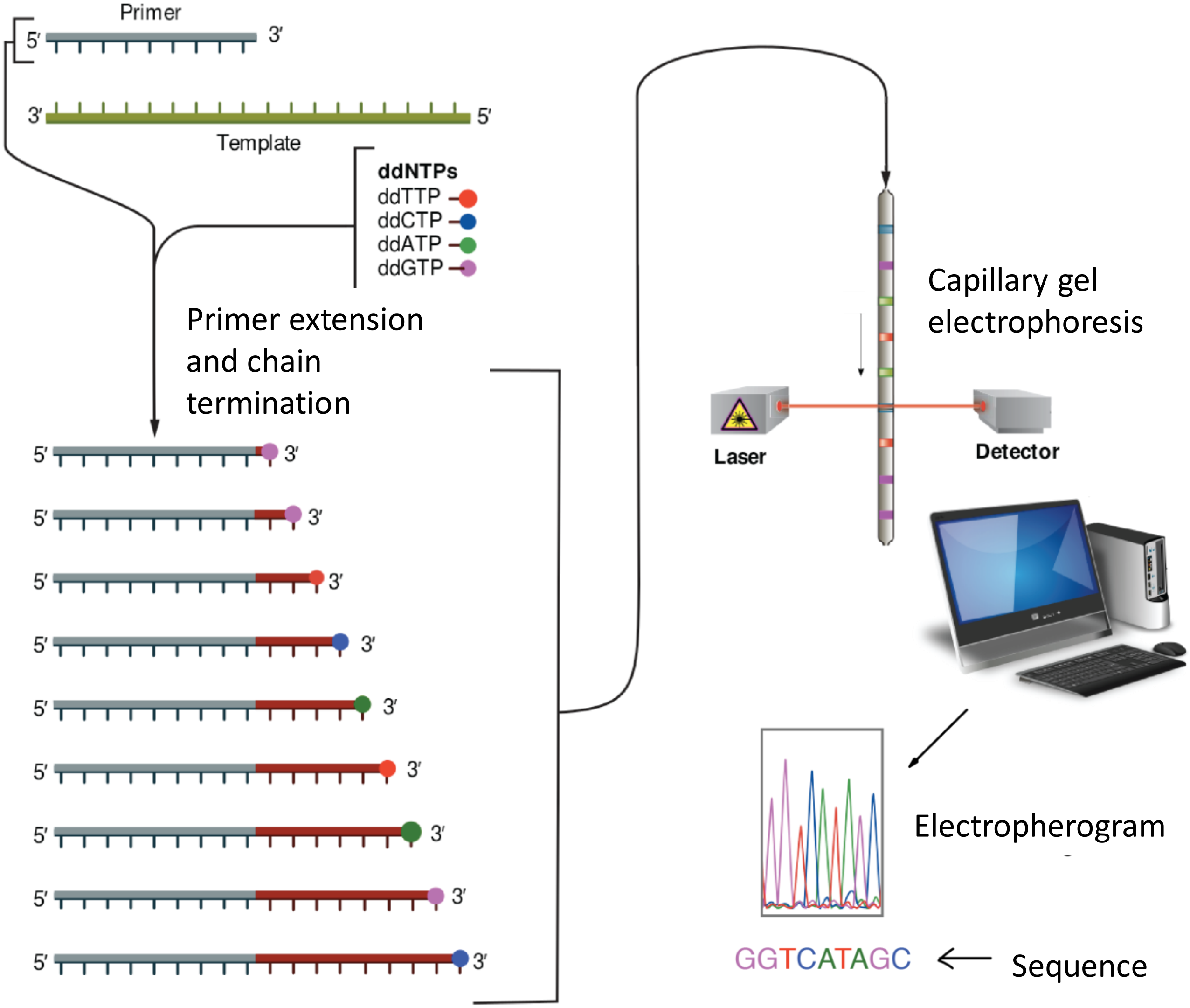
Capillary avoids interference of adjacent lanes on a gel!

(figures modified from wikipedia.org)

Sanger – applications today

- confirmation of plasmids (cloning)
- Analysis of PCR products
- Genotyping (different markers)
- Confirmation of NGS results

QUESTION: How does it work?



Primer extension and chain termination

Capillary gel electrophoresis

Electropherogram

Sequence

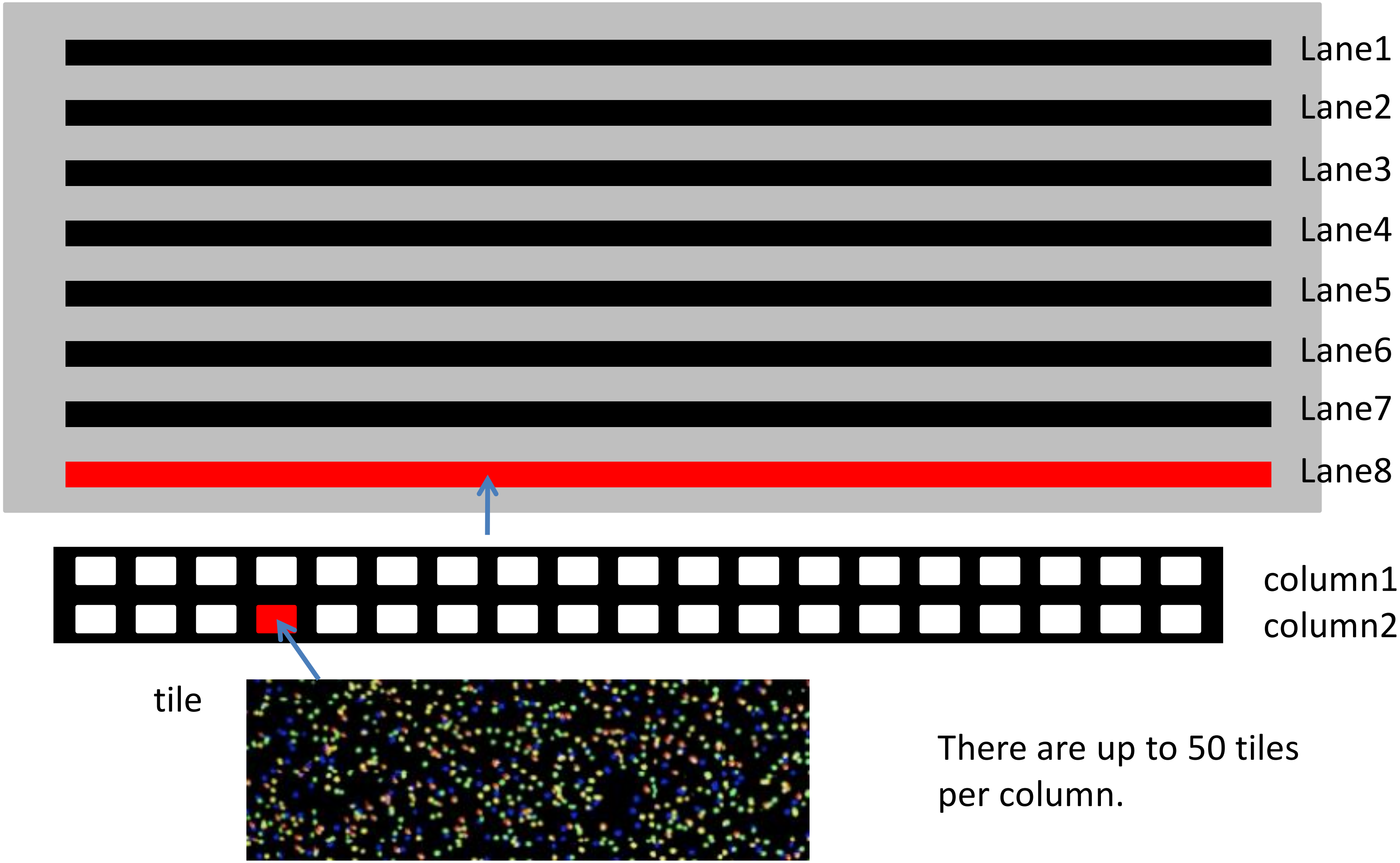
(figures modified from wikipedia.org)

NGS

Roche 454 pyrosequencing and Illumina sequencing technologies are described here:

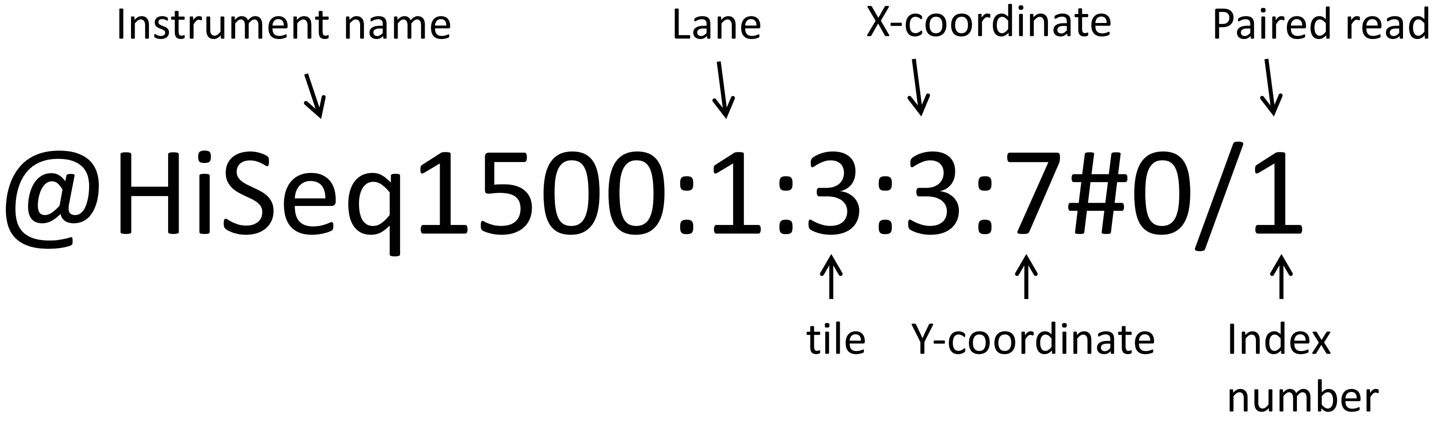
https://www.slideshare.net/PaoloDametto/new-generation-sequencing-technologies

Illumina – flow cell layout



There are up to 50 tiles per column.

Illumina – Read ID



Instrument name

Lane

X-coordinate

Paired read

@HiSeq1500:1:3:3:7#0/1

tile

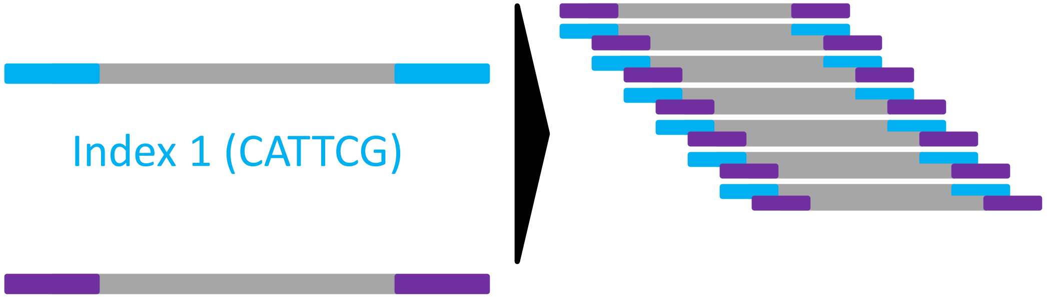
Y-coordinate

Index number

Illumina - multiplexing

Library preparation

Pool



Sequence

CATTCGACGACGAT
CATTCGGATTTCGA
AACTGAATTATTGA
CATTCGCTAGACGC
AACTGAACGATTGA
AACTGAGTCGATTG
CATTCGGATCGACA
AACTGATTGATATA
CATTCGTGCGAAGT
AACTGAGGCGATTA
AACTGATTACGAGA
CATTCGCGCGACGA
CATTCGCGATAACG

Demutliplex

CATTCGACGACGAT
CATTCGGATTTCGA
CATTCGCTAGACGC
CATTCGGATCGACA
CATTCGTGCGAAGT
CATTCGCGCGACGA
CATTCGCGATAACG

AACTGAATTATTGA
AACTGAACGATTGA
AACTGAGTCGATTG
AACTGATTGATATA
AACTGAGGCGATTA
AACTGATTACGAGA

Illumina – sequencing modi

type

- SE = single end
- PE = paired-end
- MP = mate pair

read length

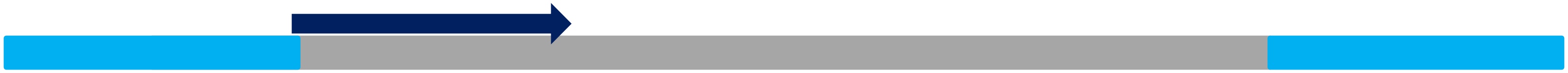
- 32nt, 50nt, 75nt, 100nt, 150nt, 250nt, 300nt

examples

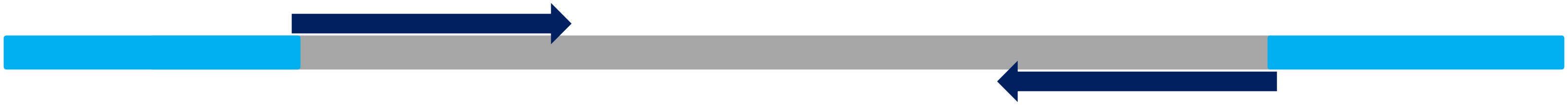
- 2x250nt PE, 2x100nt MP, 1x100nt SE

Illumina – sequencing modi

- Single end (SE):



- Paired-end (PE):

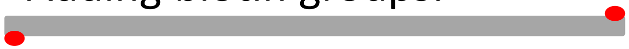


Illumina – sequencing modi

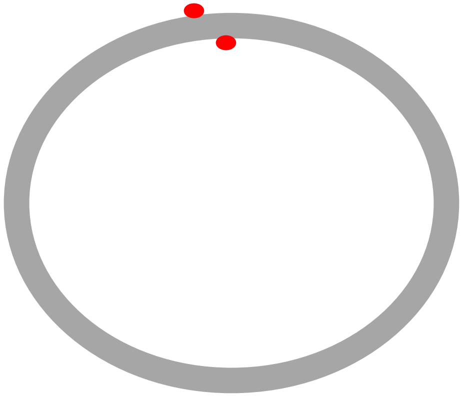
- Mate pair (MP):

Fragmentation of DNA:

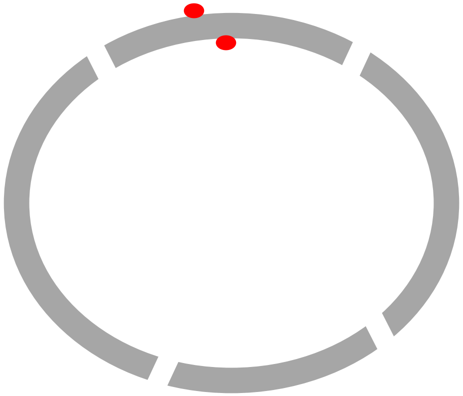
Adding biotin groups:



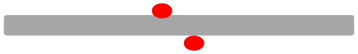
Circularization:



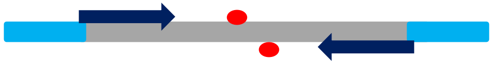
Fragmentation:



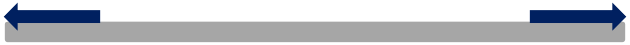
Enrichment of biotinylated fragments:



Sequencing as paired-end:



Result:

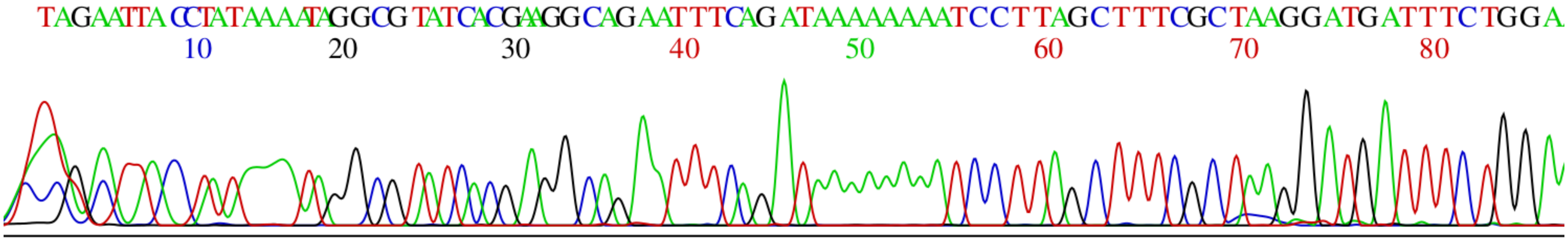


File formats of sequencing results

- TRACE
- FASTA / QUALA
- FASTQ
- SAM/BAM
- HTML

TRACE (.abi / .ab1)

- Original result file of ABI basecaller (Sanger sequencing)
- Contains only one read per file



FASTA

- There are two types of lines: header and sequence
- Header line starts with ‘>’
- Header can contain name and information about sequence
- Example:

```
>seq1 len=5
ACGTA
>seq2 len=10
ACGTA
ACGTA
>seq3 len=1
A
```

QUALA

- There are two types of lines: header and quality
- Header line starts with ‘>’, can contain name and information about sequence and one entry corresponds to a FASTA file entry
- Example:

```
>seq1 len=5
10 11 12 8 6
>seq2 len=10
10 11 12 11 11
10 10 10 6 4
>seq3 len=1
15
```

Phred-Score

- Negative logarithm of the error probability for given position in read
- Multiplication by 10 to avoid floats

Phred quality score	Error probability	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

FASTQ

- Standard format for sequences with associated quality information
- Four lines per entry:
  1. Header starts with @ (title + description)
  2. Sequence
  3. + (optional repetition of header)
  4. Quality (phred encoded in ASCII character)
- Example:

```
@seq1
ACGTACGTACGT
+
“”?CB””:DC”
```

SAM/BAM

- SAM = Sequence Alignment/Map format
- BAM = Binary version of SAM file
- Another way to store read information: contains information from FASTA and FASTQ file (reads mapped to reference)

HTML

- HTML = Hyper Text Markup Language
- Structured file format to store information
- Output format of summary produced by some tools
- Platform-independent way to combine text and figures

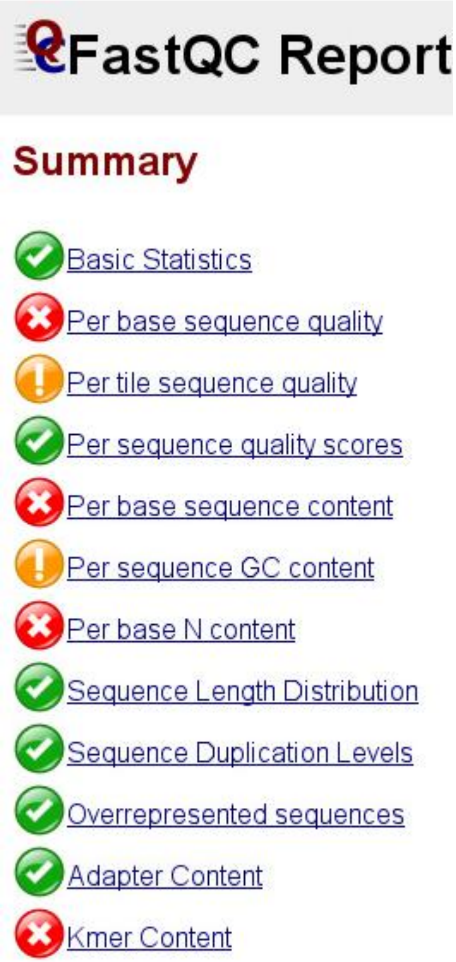
Processing of sequencing information

- Quality control (fastQC)
- Trimming (trimmomatic)
- Storage (Sequence Read Archive)
  - gzip
  - md5sum
  - filezilla

EXERCISE

- Run fastQC on your data!

fastQC



FastQC Report

Summary

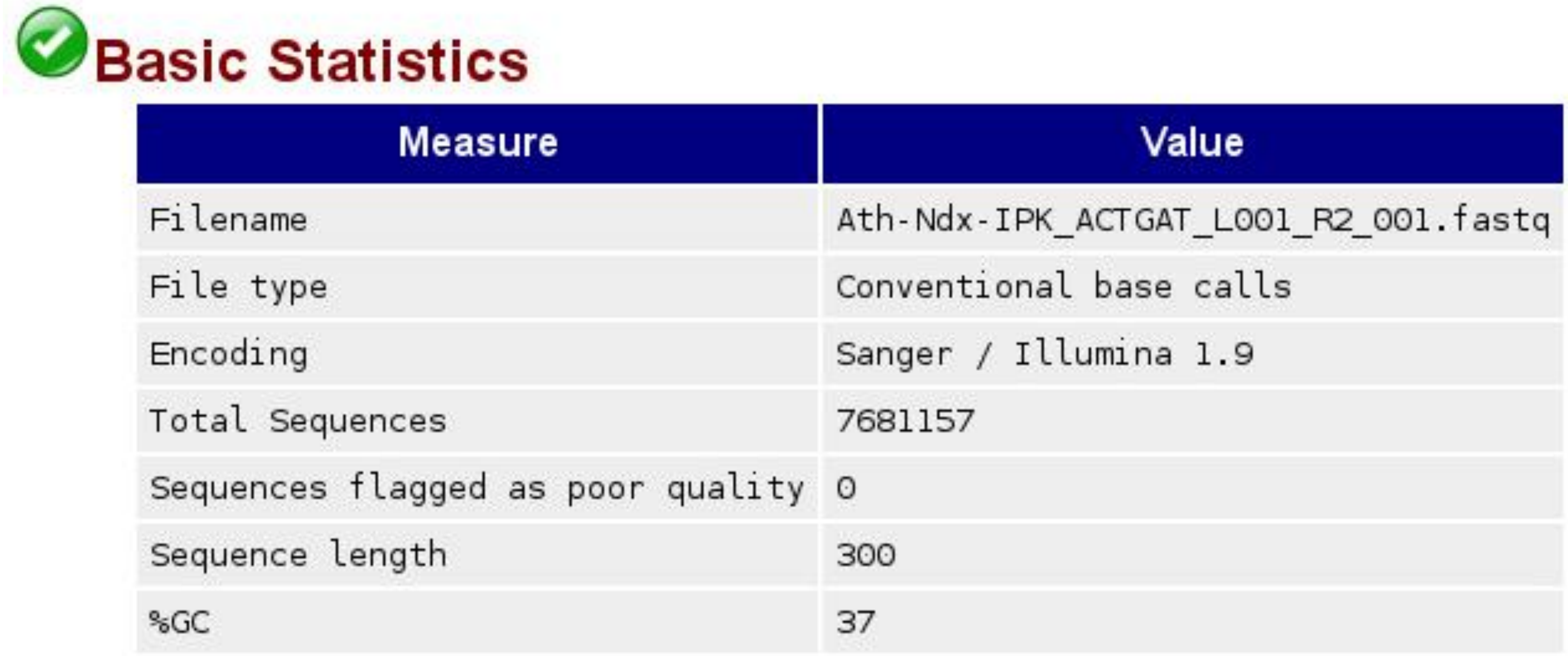
- ✓ Basic Statistics
- ✗ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✗ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

Different characteristics of the sequencing data are checked.

Results are classified as good (green), ok (orange) or bad/failure (red).

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

fastQC – basic information

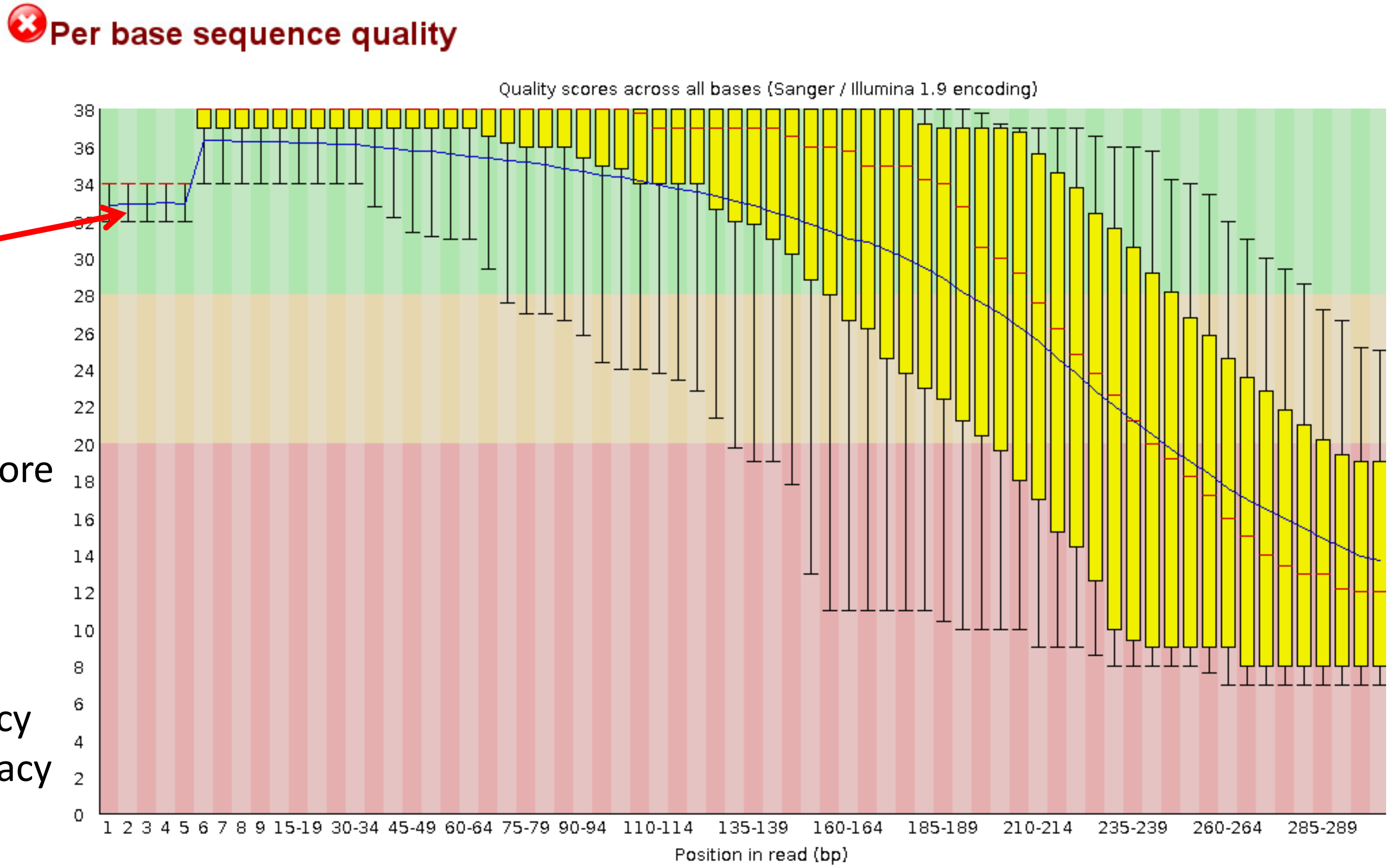


✓ Basic Statistics

Measure	Value
Filename	Ath-Ndx-IPK_ACTGAT_L001_R2_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	7681157
Sequences flagged as poor quality	0
Sequence length	300
%GC	37

fastQC – per base sequence quality

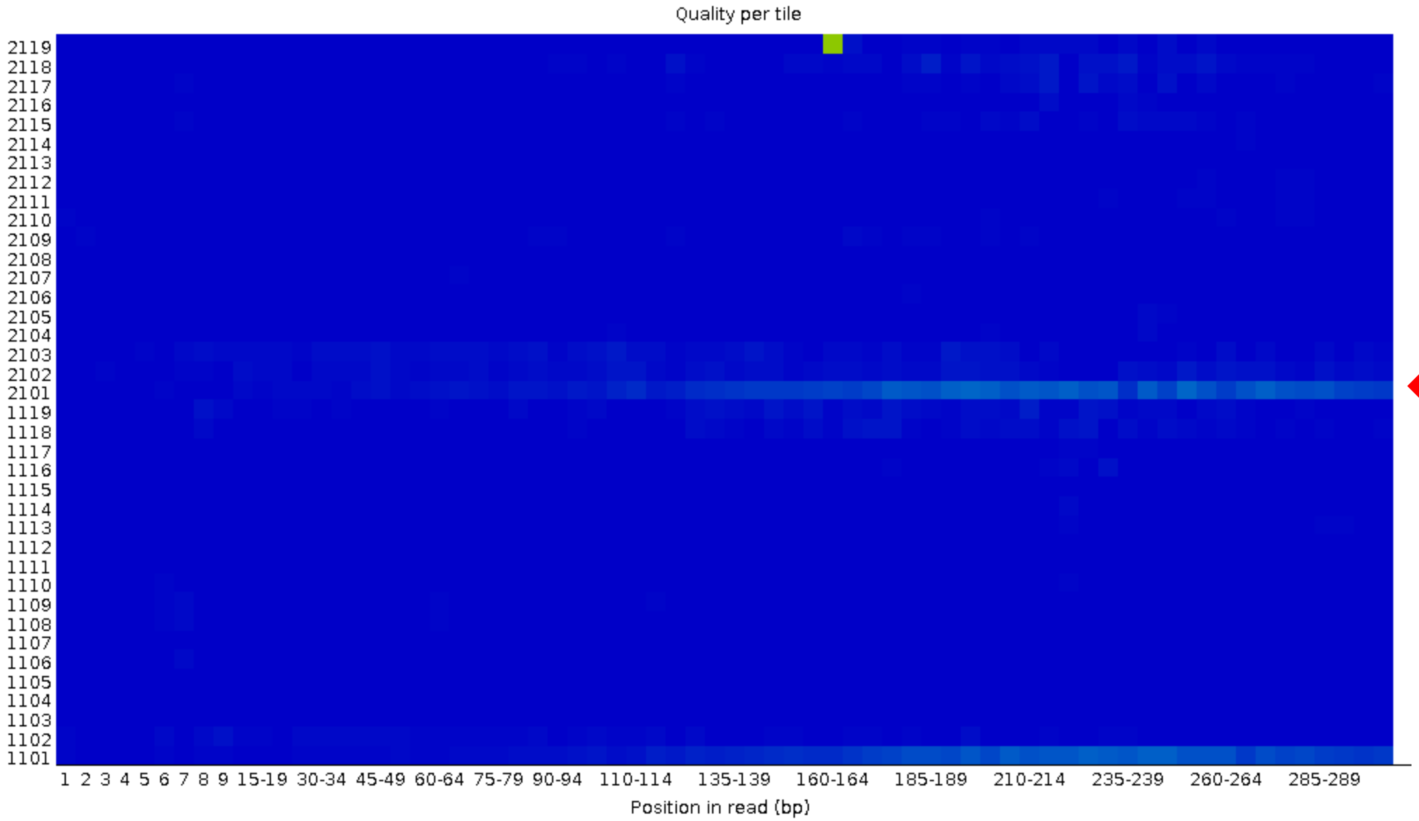
Low quality at read start is artifact caused by illumina software.



Quality value is phred score (-10* log10 of error probability):
Phred10 = 90% accuracy
Phred20 = 99% accuracy
Phred30 = 99.9% accuracy
Phred40 = 99.99% accuracy

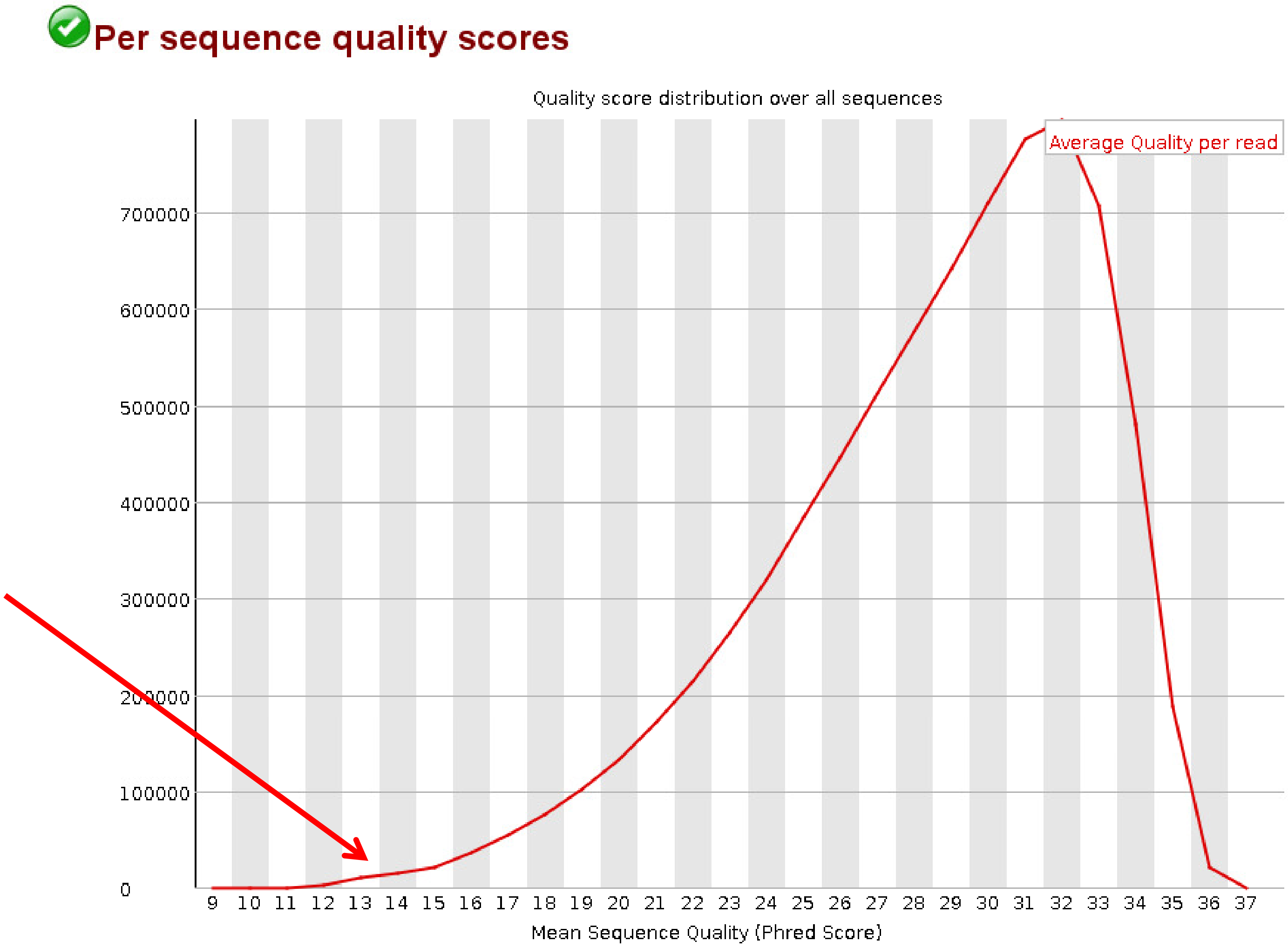
fastQC – per tile sequence quality

Per tile sequence quality



Low quality tile could be due to dust on the flow cell.

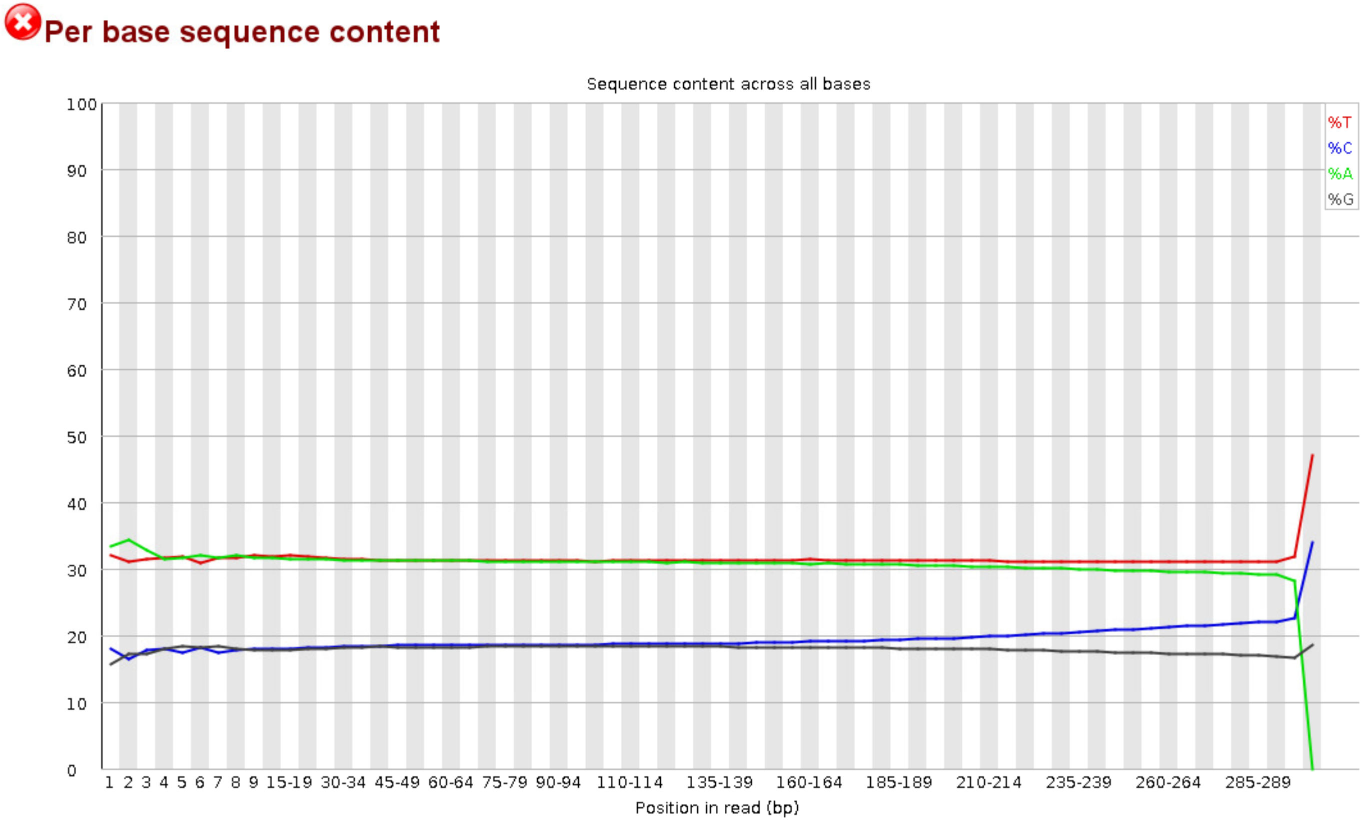
fastQC – per sequence quality score



✓ Per sequence quality scores

Low quality reads
should be removed

fastQC – per base sequence content

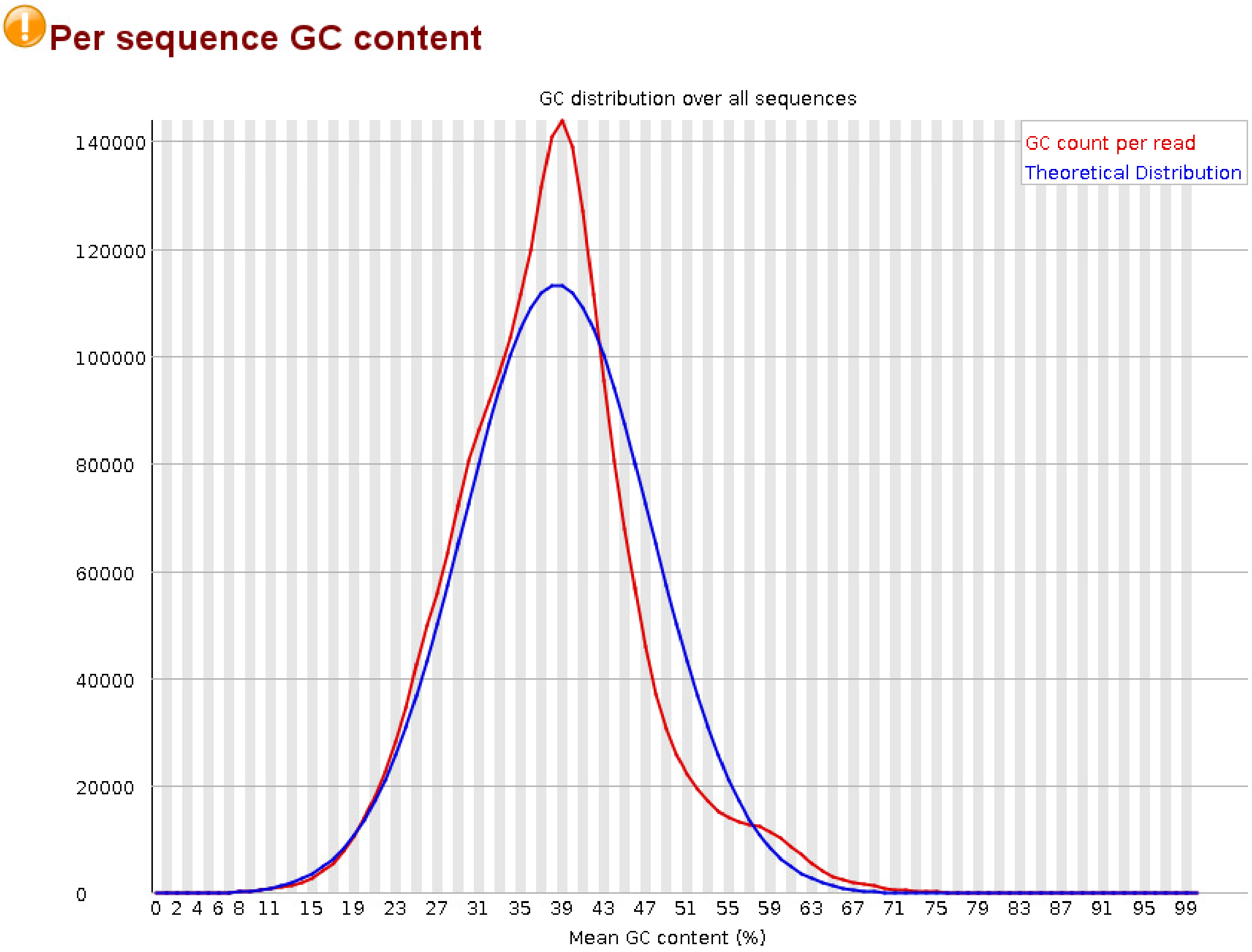


Per base sequence content

Differences at the read starts are caused by non-random fragmentation of DNA.

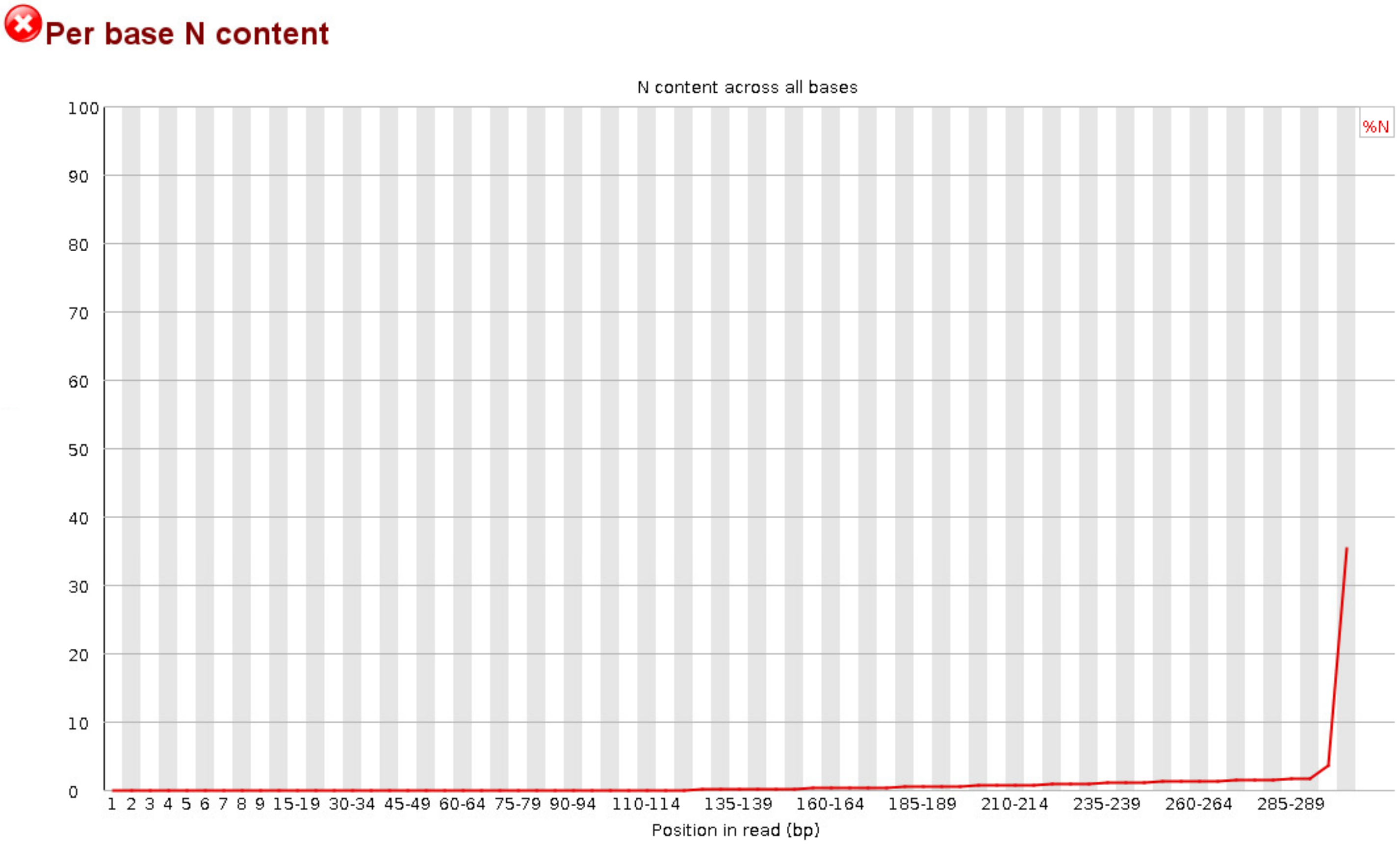
Values of complementary bases should match each other (A=T and G=C).

fastQC – per sequence GC content



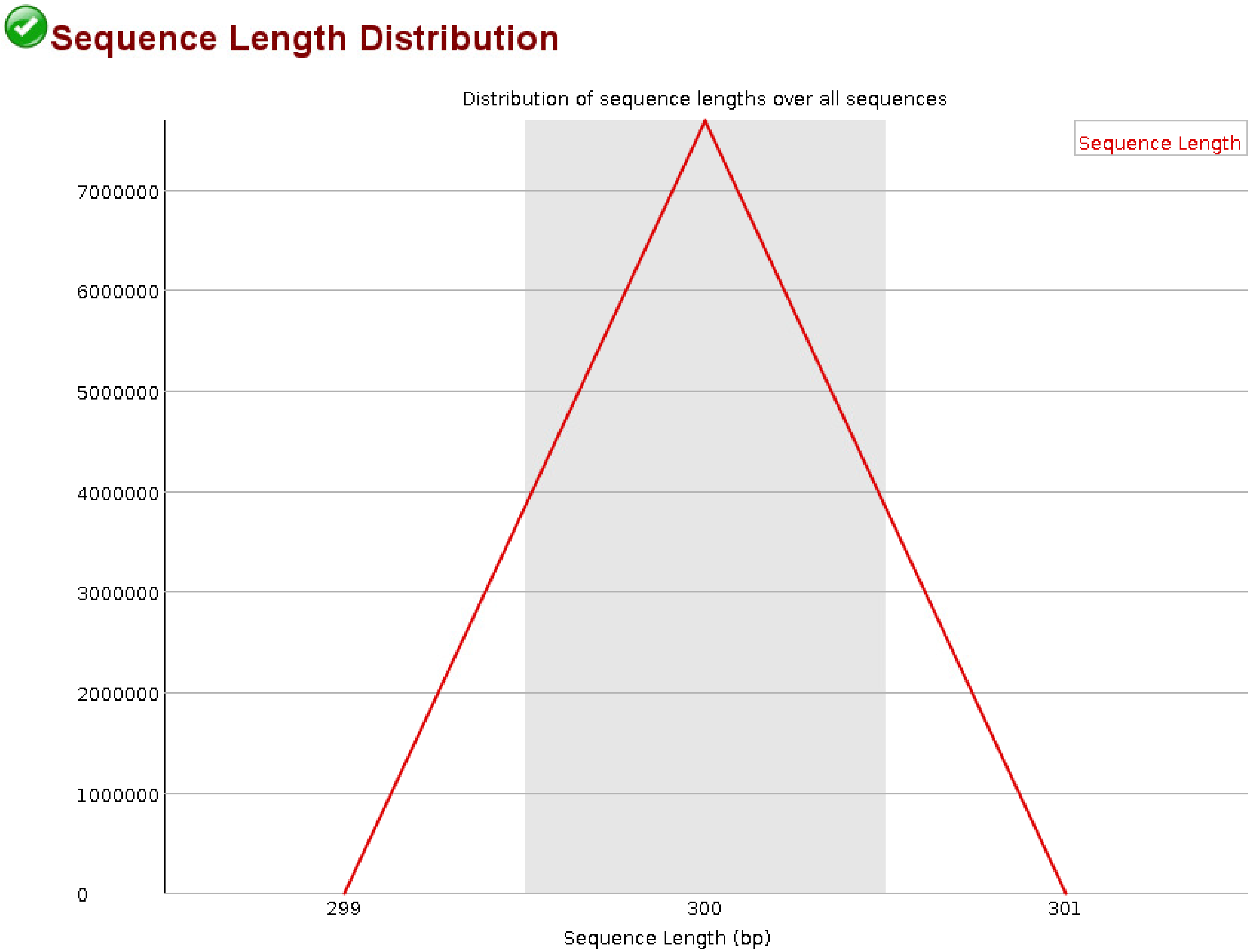
Per sequence GC content

fastQC – per base N content



All Ns should be removed by trimming.

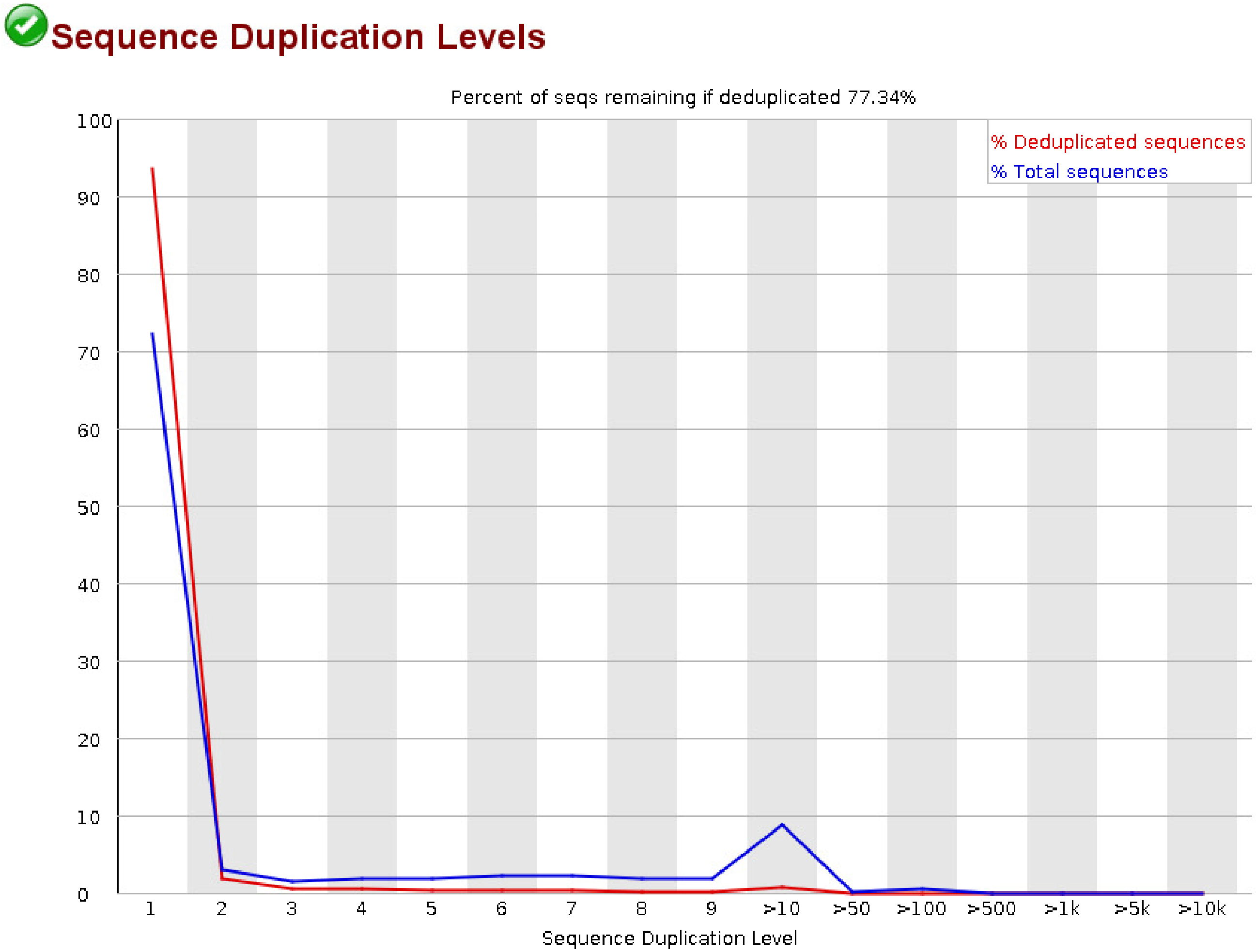
fastQC – sequence length distribution



✓ Sequence Length Distribution

All illumina reads should have the same length (at least before trimming).

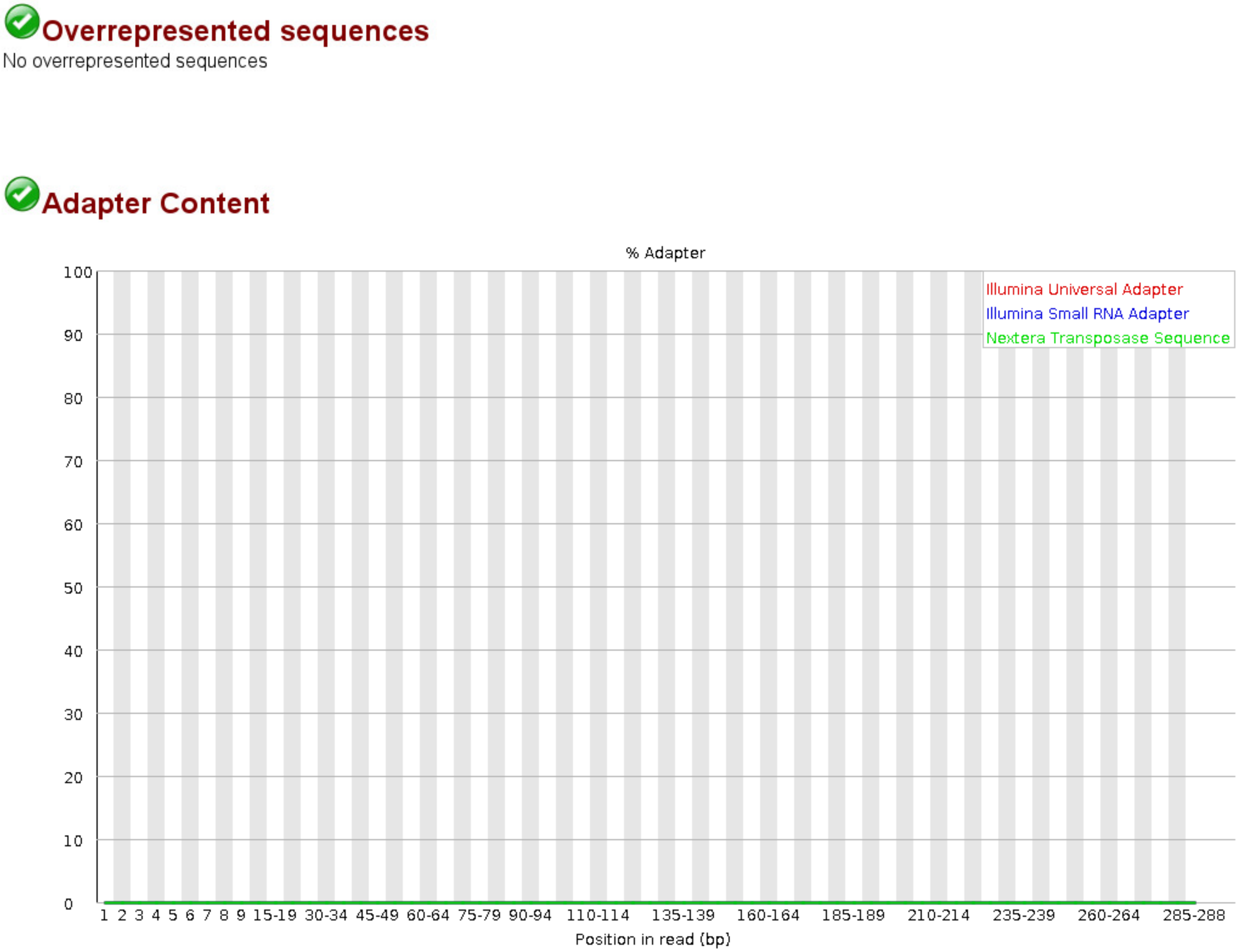
fastQC – sequence duplication levels



Duplicated sequences are reads starting and ending at the same point.

Such sequences are caused by PCR amplification of DNA fragments prior to sequencing.

fastQC – adapter contaminations



✓ Overrepresented sequences

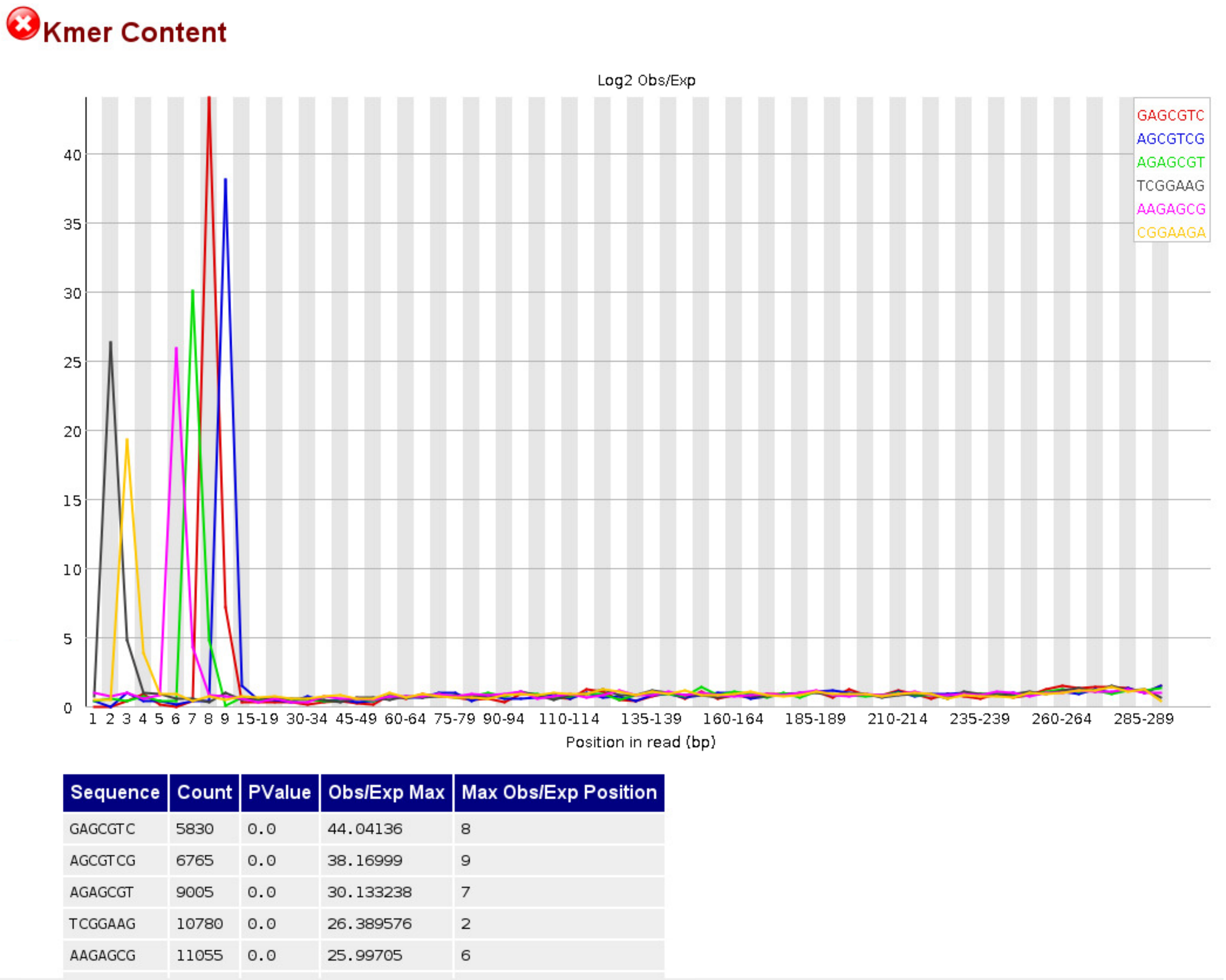
No overrepresented sequences

✓ Adapter Content

Overrepresented sequences could indicate the presence of adapter sequences in the reads.

Adapters need to be trimmed prior to assembly to avoid connections between reads via adapter sequences.

fastQC – kmer content



Kmer Content

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GAGCGTC	5830	0.0	44.04136	8
AGCGTCG	6765	0.0	38.16999	9
AGAGCGT	9005	0.0	30.133238	7
TCGGAAG	10780	0.0	26.389576	2
AAGAGCG	11055	0.0	25.99705	6

Overrepresented k-mers could indicate adapter fragments or other contaminations and artifacts.

Trimmomatic

A flexible read trimming tool for Illumina NGS data: http://www.usadellab.org/cms/?page=trimmomatic

Trimmomatic - usage

- Paired end:

```
java –jar trimmomatic-0.36.jar PE input_fw.fq input_rv.fq
out_fw.paired.fq out_fw.unpaired.fq out.rv.paired.fq
out.rv.unpaired.fq ILLUMINACLIP:TrueSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

- Single end:

```
java –jar trimmomatic-0.36.jar SE input.fq out.fq
ILLUMINACLIP:TrueSeq3-SE.fa:2:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Multi line command

```
java –jar trimmomatic-0.36.jar \
PE \
input_fw.fq input_rv.fq \
out_fw.paired.fq out_fw.unpaired.fq \
out.rv.paired.fq out.rv.unpaired.fq \
ILLUMINACLIP:TrueSeq3-PE.fa:2:30:10 \
LEADING:3 TRAILING:3 \
SLIDINGWINDOW:4:15 MINLEN:36
```

Trimmomatic –paired end

- java –jar trimmomatic-0.36.jar … call tool
- SE … select trimming modus
- -phred33|-phred64 … specify quality encoding
- input_fw.fq … input file with forward reads (mate1)
- input_rv.fq … input file with reverse reads (mate2)
- out_fw.paired.fq … paired fw reads (after trimming)
- out_fw.unpaired.fq … unpaired fw reads (after trimming)
- out_rv.paired.fq … paired rv reads (after trimming)
- out_rv.unpaired.fq … unpaired rv reads (after trimming)
- ILLUMINACLIP:<FILENAME> …. multiple FASTA file contains adapter sequences for clipping

Trimmomatic – paired end II

- LEADING:<INT> … number of leading nucleotides to remove
- TRAILING:<INT> … number of nucleotides to remove at read end
- SLIDINGWINDOW:<WIN_SIZE>:<QUAL_CUTOFF> … specifying a sliding window to trim read at low coverage position
- MINLEN:<INT> … minimal length of read after trimming to prevent discarding
- TOPHRED33|TOPHRED64 …. converts read quality scores into phred33 or phred64 scale

Trimmomatic – paired end III

- Autodetection of used phred score
- Basename of input and outup file
- Trimming summary report at end of process

EXERCISE

- Run Trimmomatic on your data!

Data storage - gzip

- FASTQ files consume much space on device
- File size can be reduced significantly by different compressions
- NGS data should always be processed in compressed format
- Almost all tools support compressed data input

- Compression of read data via gzip:

```
$ gzip <FASTQ_FILE>
```

- Decrompression:

```
$ gunzip <FASTQ_FILE>
```

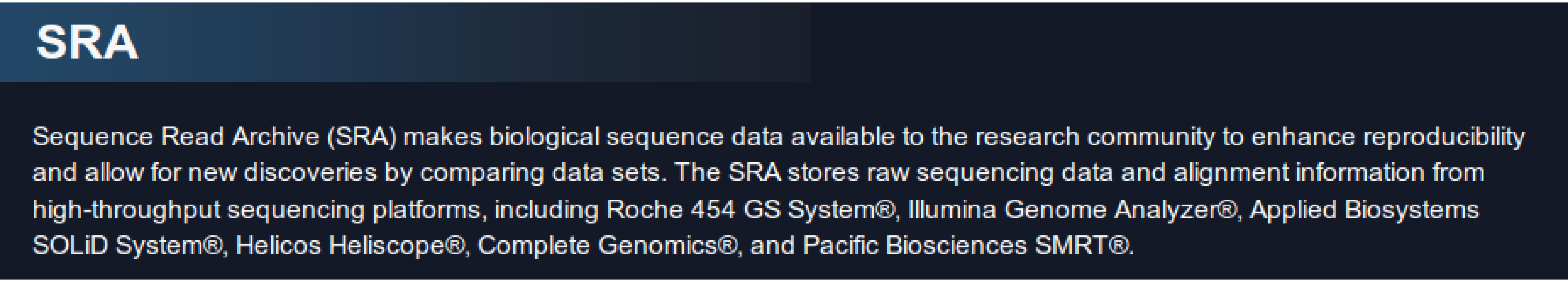
EXERCISE

- Compress all your data via gzip!

QUESTIONS

- How large is the size difference between compressed and uncompressed files?
- How is it possible to extract data from .gz files?
- How is it possible to make sequencing data available to the whole science community?

Sequence Read Archive (SRA)



SRA

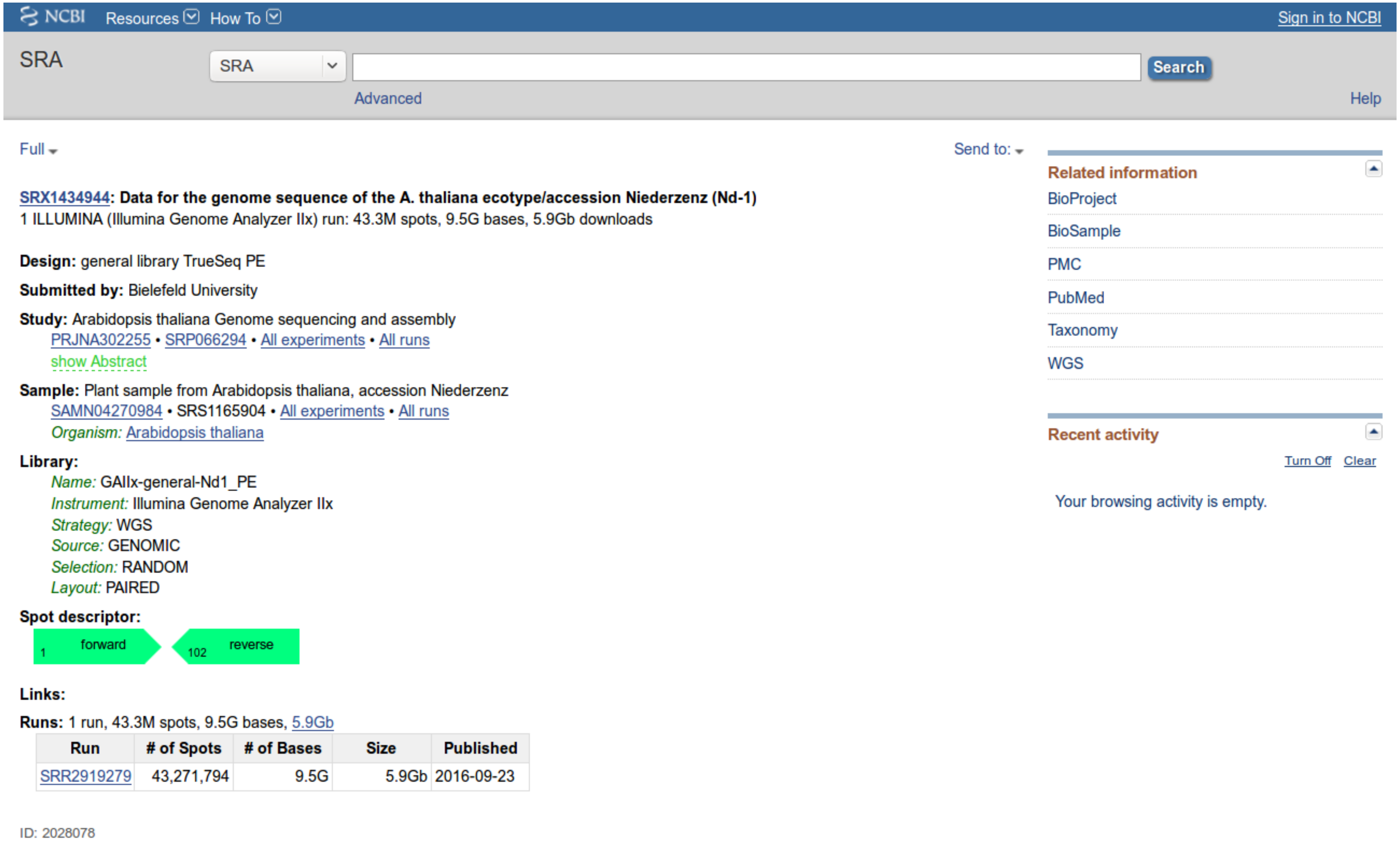
Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Data submission:

1) Construct entry with meta information
2) Submit corresponding data files

Example:
https://www.ncbi.nlm.nih.gov/sra/SRX1434944

SRA - Example: SRX1434944



SRX1434944: Data for the genome sequence of the A. thaliana ecotype/accession Niederzenz (Nd-1)
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 43.3M spots, 9.5G bases, 5.9Gb downloads

Design: general library TrueSeq PE

Submitted by: Bielefeld University

Study: Arabidopsis thaliana Genome sequencing and assembly
PRJNA302255 • SRP066294 • All experiments • All runs
show Abstract

Sample: Plant sample from Arabidopsis thaliana, accession Niederzenz
SAMN04270984 • SRS1165904 • All experiments • All runs
Organism: Arabidopsis thaliana

Library:
Name: GAIIx-general-Nd1_PE
Instrument: Illumina Genome Analyzer IIx
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:

Links:

Runs: 1 run, 43.3M spots, 9.5G bases, 5.9Gb

Run	# of Spots	# of Bases	Size	Published
SRR2919279	43,271,794	9.5G	5.9Gb	2016-09-23

ID: 2028078

SRA - md5sum

- specific for one file
- used to compare files
- needs to be submitted to the SRA to check completeness of transferred data files
- consists of 32 positions of 0-9 and a-f (hexadecimal system)
- Linux offers a function for md5sum calculation:

```
$ md5sum <FILENAME>
```

QUESTION

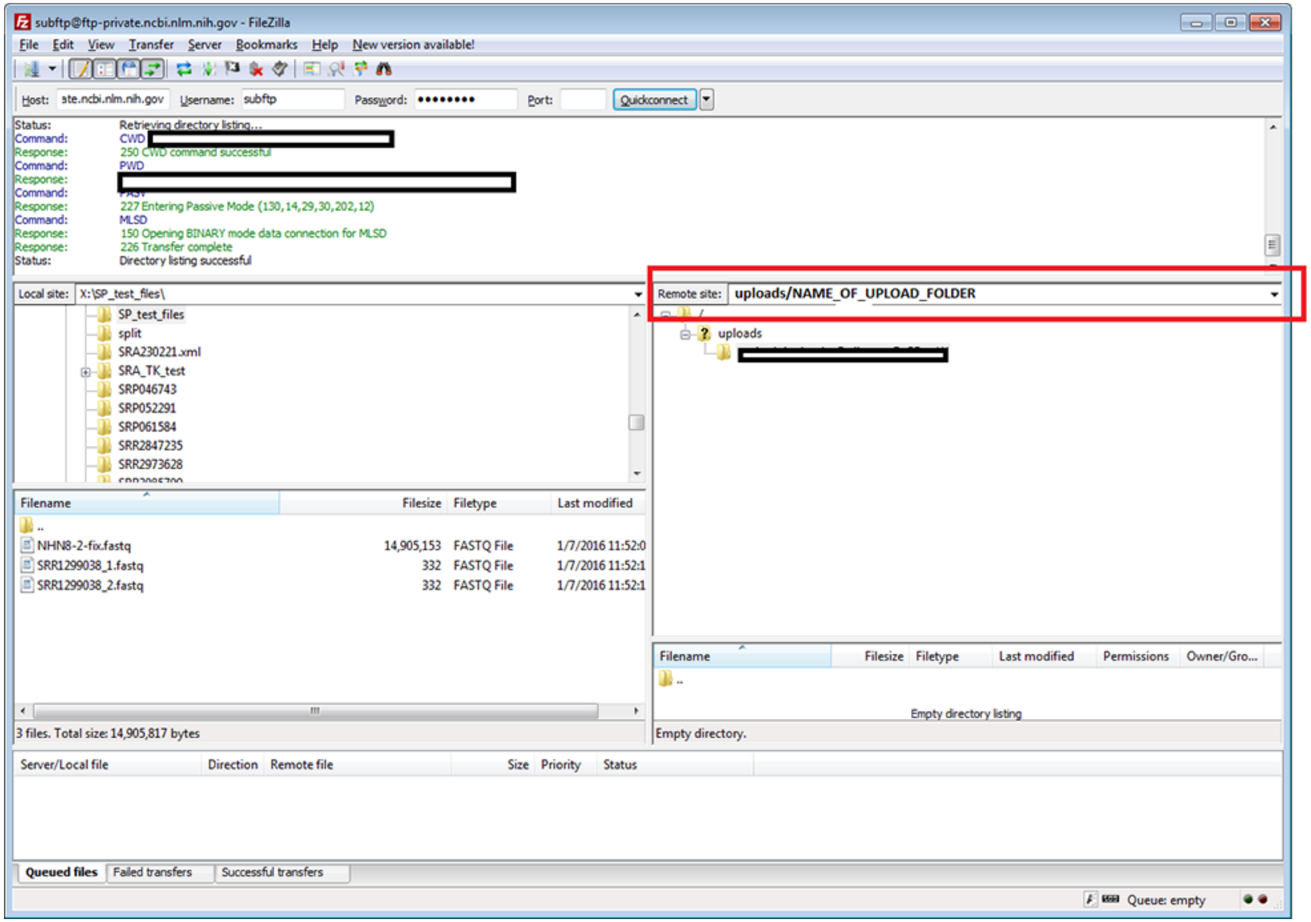
- What are the md5sums of your trimmed paired reads FASTQ files?

QUESTION

- What are the md5sums of your trimmed paired reads FASTQ files?
  - Fw: 0f91fa93d00849d9dfd6fa3d66a184e1
  - Rv: 84df759a82e223f6fb7c9faa85b2e60d

SRA – FileZilla for data upload

- Transfer of data files to the SRA via FTP or FileZilla



subftp@ftp-private.ncbi.nlm.nih.gov - FileZilla

File Edit View Transfer Server Bookmarks Help New version available!

Host: ate.ncbi.nlm.nih.gov Username: subftp Password: •••••••• Port:

```
Status:     Retrieving directory listing...
Command:    CWD [redacted]
Response:   250 CWD command successful
Command:    PWD
Response:   [redacted]
Command:
Response:   227 Entering Passive Mode (130,14,29,30,202,12)
Command:    MLSD
Response:   150 Opening BINARY mode data connection for MLSD
Response:   226 Transfer complete
Status:     Directory listing successful
```

Local site: X:\SP_test_files\

- SP_test_files
- split
- SRA230221.xml
- SRA_TK_test
- SRP046743
- SRP052291
- SRP061584
- SRR2847235
- SRR2973628

Remote site: uploads/NAME_OF_UPLOAD_FOLDER

- /
  - uploads
    - [redacted]

Filename	Filesize	Filetype	Last modified
..			
NHN8-2-fix.fastq	14,905,153	FASTQ File	1/7/2016 11:52:0
SRR1299038_1.fastq	332	FASTQ File	1/7/2016 11:52:1
SRR1299038_2.fastq	332	FASTQ File	1/7/2016 11:52:1

3 files. Total size: 14,905,817 bytes

Filename	Filesize	Filetype	Last modified	Permissions	Owner/Gro...
..					

Empty directory listing

Empty directory.

Server/Local file	Direction	Remote file	Size	Priority	Status

SRA – download data via fastq-dump

- Download data files from the SRA via command line
- Change to data directory (huge temp files will be stored in working directory!!)
- Paired-end reads will be placed in two files and compressed in gzip format
- Usage:

```
$ fastq-dump - -split-files - -gzip <SRR_ID>
```

- Example:

```
 $ fastq-dump - -split-files - -gzip SRR3340908
```

QUESTION

- What do you know about SRR3340908?
- How can you get more information?
- Are there more entries related to the same project?